Advances in structure and small molecule docking predictions for crystallized G-Protein coupled receptors

Dahlia Anne Goldfeld

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Advances in structure and small molecule docking predictions for crystallized G-Protein coupled receptors

Dahlia Anne Goldfeld

This dissertation discusses two main aspects of protein-ligand interaction for G-Protein coupled receptors: structure predictions of the flexible loop domains and docking into these receptors. The prediction of loop structure has been long worked on in the context of native, globular proteins. In this work it is extended to transmembrane proteins, which requires an explicit integration of the lipid bilayer into the loop prediction calculation. In the initial work, this new approach to loop prediction yields highly accurate 3-dimensional structures of the intra and intercellular loops of four G-protein coupled receptors—the A2A adenosine, bovine rhodopsin, β1 and β2 adronergic receptors. For these cases, the loops were predicted in the context of a completely native crystal structure. In subsequent work the approach was extended to work on perturbed cases, where all loops and tails were removed, and side chains near the loop being predicted were in nonnative conformations. Lastly, a full homology model of the β2 adronergic receptor was successfully built from the β1 adronegric receptor as its template. Work on docking into these receptors focuses on the kappa opioid receptor. Known antagonist binders are discriminated from a set of decoy nonbinders via docking calculations. Two new terms were added to the scoring function, WScore to achieve this, based on a detailed molecular understanding of how the receptor works.

Table of Content

List of Tables

List of Figures

Acknowledgments

I would like to use this space to thank and acknowledge several people who have been integral to me in completing my Ph.D. First and foremost I need to thank Richard Friesner for his invaluable support and mentorship throughout the years. He has been a fantastic advisor, both scientifically and personally. He never ceases to have new ideas to attack challenging problems, and he always remains enthusiastic and positive in the face of temporary failures. I would also like to acknowledge the other members of my committee, all of whom have truly had a positive impact on me: Angelo Cacciuto, for his guidance and support, and particularly for helping me transfer to the graduate school; Barry Honig, for his counsel and willingness to talk to me about future plans, as well as superior editing skills; Bruce Berne, for his generosity with time to discuss the scientific world with me; and lastly James Leighton, for bringing us to the kappa opioid receptor project and hopefully a fruitful collaboration as we continue to understand the protein.

Many thanks must also be given to coworkers at Schrödinger—Robert Murphy, Kai Zhu, and Thijs Beuming—each of whom contributed to the success of my three main papers over the last three and a half years. Robert Murphy remains the best scientific programmer with whom I have ever had the privilege of working, and I am grateful for his assiduousness and support. Woody Sherman and John Shelley also provided very helpful advice. At Columbia, I owe thanks to Edward Miller, with whom I have worked closely on the protein local optimization program, as well as the other members, past and present, of the Friesner group: Shulu Feng, Michelle Hall, Steven Jerome, Jianing Li, Peilin Liao, Colleen Murrett, Andrew Weissman, and Jing Zhang. Lastly, I would like to thank Cal Lobel for his hard work and dedication to our cluster. Without him, much less would have ever been accomplished.

I dedicate this thesis to my family and to tea farmers worldwide.

Chapter 1

Introduction

1.1 Protein-ligand interaction and drug discovery: the importance of protein structure and binding free energy calculations

Proteins are cells' workhorses, performing a wide array of biological tasks and functions (1). Proteins interact with ligands, other proteins, and surfaces, controlled by various intermolecular forces within and outside the binding site. The ability to interact broadly—one protein can bind a wide array of proteins, nucleic acids, and small molecules, as well as adhere to various surfaces—comes from the structural reorganization of the binding pocket that occurs upon binding. The intrinsic binding site flexibility is essential for our own endogenous proteins and small molecules, as well as pharmaceuticals that exploit a receptor, either by activating (an agonist) or deactivating (an antagonist) it, for medical purposes. The binding site is made up of amino acids that constitute both core and loop regions of a protein. It is the precise sequence and orientation of these residues that allow for substrate recognition and fit inside a protein. Accurate knowledge of the shape of the active site (including loops) is thus very important.

The geometry of a binding molecule within a receptor depends on the docking program used, and binding free energy estimates (which discriminate the poses from one another, separates active from non-active ligands, and rank order binding affinities) depend on the potential function used for the calculation. The primary types of potentials for such calculations are empirical (2), knowledge based (3, 4), and first principle based (5, 6). Empirical scoring functions try to estimate the contributions of various interaction terms (hydrophobic, hydrophilic, rotatable bonds, etc) to binding affinity. The coefficients of each term are optimized with multiple linear regression by fitting predicted and experimentally known affinities for a training

set of ligand-receptor complexes. Problems with empirical methods arise from the construction of the training set (how large it is, how representative it is of ligand-receptor complexes at large, and other concerns). Furthermore, because the interaction terms are interpolated to other complexes, novel molecular scaffolds cannot be found as they are not present in the training set. Knowledge based scoring functions are constructed using a sets of rules (that rely on statistical analysis of experimentally determined intermolecular close contacts of very large numbers of ligand-receptor complexes) that are turned into pseudo-potentials. Like all knowledge-based methods, success is restricted to formerly observed patterns, in this case of ligand binding, while novel binding motifs are almost definitely missed. Lastly, first principle based scoring functions estimate the van der Waals and electrostatic interactions between ligand atoms and the part of the receptors that they occupy. This is typically accomplished by calculating gas phase molecular mechanics contributions and combining it with solvation free energy, which is challenging. Unfortunately, the accuracy of first principle based methods (also called force fields) is generally limited to ranking ligands based on relative binding affinity. There are also hybrid methods which are based on force fields and add new knowledge based or empirically derived terms, but this is very much a problem in flux.

A natural extension of being able to dock a molecule to a receptor and estimate its binding affinity is to modify *in silico*, either the ligand or the active site, to increase affinity. This is the fundamental idea behind rational drug design (7, 8). First a virtual ligand screen is done on a protein receptor of interest to provide a prospective lead, which is verified experimentally to have biological activity. The molecule is then optimized. Of course, this procedure is predicated on the idea that the correct binding mode and affinities of a wide array of ligands can be accurately captured. While there are many methods and programs available to approach this

problem (GLIDE (9-13), DOCK (14), FLEXx (15), GOLD (16), etc), the docking and scoring problem is central to computer aided drug design is far from solved. It is an inherently difficult problem, requiring enough conformational sampling of both the ligand and binding site as well as accurate energy evaluations in spite of rugged binding energy landscapes.

Complicating matters further, for many receptors, we only have one crystal structure, lending experimental insight into only one binding pocket conformation and binding mode. It is widely agreed upon that ensemble docking, in which there are at least three or four crystal structures of a given receptor, yields much more accurate docking results. Alternative active site conformations can be found using long molecular dynamics simulations and other techniques, but the quality of the resultant structures can be questionable. Beyond this, despite great advances in x-ray crystallography techniques, we do not have even one crystal structure for a tremendous number of pharmaceutically useful proteins. In this situation, the best approach to working with a protein *in silico* is to create a homology model of it. Homology models are built by modeling the structure of the protein of interest after a homologous protein whose structure is known. For extremely similar proteins, this approach can yield a very accurate model, but much of the time homology models are quite different from the true structure. Typically, the most accurate regions of a homology model will be the parts that are closest in sequence to the template structure and the parts that are the most geometrically constrained—namely helices and beta strands. The most inaccurate areas in homology models tend to be the flexible regions such as loops. Finding different configurations of the active site starting with a homology model is even more difficult and probably wrong the majority of the time. Models further distort the underlying binding energy landscape of the active site, making binding predictions for them exponentially more ambitious and inaccurate. Thus, the best of the computer aided drug

discovery process is currently limited to x-ray crystal structures. This will continue to be true until computational determination of protein structure becomes a more robust and reliable method.

The lack of structures extend across all protein classes but are particularly notable for membrane receptors, special proteins that recognize, decode, and transduce extracellular signals (17). They are also notoriously difficult to crystallize (18, 19). Although 20% of the human genome are membrane proteins, less than 2% of solved protein structures are of membrane proteins (19). They also tend to be low resolution structures. It is very difficult to produce membrane proteins in active, folded forms in sufficient yields, because selecting appropriate membrane-mimicking environments that support their function and stability is very tricky (20). The lipid bilayer, or membrane, environment in which they reside is interesting, composed of two layers of lipid molecules. The membrane itself forms a thin, flat sheet that is a cell's barrier. Mostly composed of phospholipids, which have a hydrophilic head and hydrophobic tail, they are arranged in a two-layered sheet called the bilayer, with the tails pointing inward. Other lipids and cholesterol also make up these membranes.

The detergents commonly used for membrane protein crystallization preparation are also disruptive to the protein's stability, which has lead to the use of phospholipid bilayers and phospholipid nanodiscs as membrane mimetics to overcomes these problems. Improving membrane protein crystallization is a topic of active research and has had a huge influence in the study of G-protein coupled receptors.

G-protein coupled receptors (GPCRs) are the largest family of transmembrane proteins. They mediate responses to a vast number and variety of bioactive molecules, making them crucial to basic physiological functions such as neurotransmission, cellular metabolism, cell

growth, blood pressure regulation, and immune defense (21). They are also implicated in human pathologies ranging from tumor metastasis to Alzheimer's disease (22), making them very valuable drugs targets for drug design.

There are five families of GPCRs (23), all characterized by seven transmembrane (TM), hydrophobic alpha-helices that are connected by alternating intra and extracellular loops. The active sites of GPCRs are complex locales, where ligands can in theory interact with the TM region, loop domains, crystal waters, and the lipid bilayer, all at the same time. For most of the crystallized class A GPCRS, the TM bundle defines the main binding pocket for ligands, with extracellular loop contacts being the secondary factor (24). Accurate knowledge about the loop structures in GPCRs can thus be essential toward predicting ligand binding as well as understanding the molecular strategy of signal transduction on the intracellular side of the protein.

Unfortunately, like other transmembrane proteins, crystal structure determination of GPCRs is formidable, exemplified by the fact that until 2007, only rhodopsin in its dark state had been resolved with atomistic detail. Since 2007, however, there has been an enormous advance in crystallizing GPCRs, and at the time of writing there is at least one crystal structures for each of the 18 unique GPCRs (25). It seems reasonable to now claim that we have finally entered the era of structure-based research for GPCRs. This diversity of new structures is exciting, and we are no longer as wholly dependent on GPCR homology models (between 2000 and 2010 the number of papers in the literature focusing on homology models of GPCRs based on rhodopsin, and then the adrenergic receptors, is astounding). Nonetheless, there remain too few structures to thoroughly study GPCR structure and function and to screen potential pharmaceutically active compounds on all GPCR targets of interest, and this will remain true for many years. Indeed,

there are still no published GPCR structures of any GPCR outside of the class A receptors; there is still a ways to go.

In the absence of experimental structures, homology modeling presents an option for predicting new GPCR structures. Even restricting such models to family A receptors, this is still an extremely difficult problem. The overall sequence identity and 3-dimensional structural similarity amongst family A receptors is low, restricted to a few, key highly conserved residues scattered throughout the TM helices (26). Even between highly similar GPCRs, like the β1 and β2 adronergic receptors or M1 and M2 muscaneric receptors, the tails and loops can differ dramatically in conformation. The explosion of new GPCR templates is helpful, as the likelihood of having a better template upon which to build a homology model has increased. However, the majority of GPCR targets will not have such a close homologue that has been crystallized. Even if a good prediction of the TM domain can be obtained by aligning these key residues, the loop regions exhibit extra challenges for two primary reasons: Loops tend to be the least conserved regions amongst similar proteins and they display high conformational flexibility. This is particularly true for long loops, which are present in all GPCRs whose structures are available.

Thus, being able to computationally predict the structure of GPCR loops is very important. The first step in such an endeavor is to predict loop structures within the context of an already crystallized protein (this is the only way to validate that methods are working). Only then is it justified to move on to homology models, where loop refinement is sorely needed. To do such types of loop predictions, we utilize the Protein Local Optimization Program (PLOP) which has been developed over the past decade (27). It predicts loop regions or single side chain conformations in the context of the 3D structure of the rest of the protein (either the crystal structure or a homology model). At the present, PLOP can predict—in the context of the crystal

structure—with sub-1Å accuracy, loops ranging between 4 and 20 residues (28, 29), and it can now also predict loops up to 27 residues in length found in GPCRs (30, 31).

Both protein structure and, as discussed previously, docking predictions have two fundamental problems: sampling and scoring. With respect to PLOP, sampling the space that loops can occupy requires an algorithm that searches conformational space and generates an ensemble of candidate structures and positions. The algorithm must cleverly discretize space, or combinatorial explosion of possible loop space and docking conformations makes the problem intractable. The PLOP sampling algorithms are described in depth in the text of this thesis, and particularly in Chapter 6.

An accurate scoring function is essential to rank and select top loop candidates. For a scoring function to pick out loops that are close to the native structures, (similar to picking out ligands that are bound in the most thermodynamically stable way to a receptor) it must be able to correctly describe the free energy of the system. A scoring function is only useful if a lower energy structure is indeed closer to the global minimum of its Gibbs free energy surface than a higher energy structure. The scoring function within PLOP has been modified several times since it was initially constructed, although I did not directly work on it. It is physics based—composed of a molecular mechanics force field and a continuum solvation model based on the generalized Born model. The force field itself is a function of inter-nuclei distance and decomposed into the following terms: bond stretching angle bends, torsion, Coulomb interactions, and van der Waals interactions. On top of this, the most modern version of PLOP's scoring function contains several empirical (yet still physics rationalized) terms that directly focus on hydrophobicity and interactions of stacked aromatic rings, amongst others.

The continuum solvation model is an alternative to explicitly incorporating thousands of water molecules in an energy calculation (which is extremely time consuming and impractical). In a continuum solvation framework, the waters are treated as structureless and have a high dielectric constant (ie. 80) as compared to the protein atoms (around 4, depending on the model). This allows accurate electrostatic solvation free energy calculations.

In this thesis, we examine the problems of predicting the structure of the highly variable loop regions of GPCRs and building homology models. We also study docking into these receptors. These are crucial problems to be solved for future drug discovery efforts for GPCRs and are all key elements to the more general investigation of protein-ligand interaction. Although there is still much work to be done we have made great strides.

The GPCR loop prediction successes with PLOP were made possible by inclusion of the membrane in the calculations, as well as a new sampling algorithm called the phase space partitioning method. The former served to block regions of space where a loop could not be present, either due to steric or electrostatic reasons. The latter increased the amount of sampling done on loops. These ultimately lead to us demonstrating that loop refinement in the context of an accurate GPCR homology model is possible. The inclusion of a membrane is computationally tricky, since the complexity, including charge accuracy, of the model phospholipids are not always so accurate. Nonetheless, this is a very encouraging step forward, and new methods are being developed to further improve loop refinement in homology models.

On the docking side of the GPCR work, we used GLIDE in combination with a new scoring function, WScore to elucidate a binding mode of the Kappa opioid receptor (KOR) (32) and to separate a set of actives from accompanying decoys. To do this, the scoring function,

WScore, was modified, adding empirically derived terms to the force field (which itself is based on OPLS-AA) (33, 34) that are essential to correctly capture binding to the KOR.

In the remainder of this introduction, a brief summary of each coming chapter is given.

1.2 Successful prediction of the intra- and extracellular loops of four G-protein coupled receptors

Chapter 2 presents the results of the restoration of all crystallographically available intra- and extracellular loops of four G-protein coupled receptors (GPCRs): bovine rhodopsin (bRh), the turkey β-1 adrenergic receptor (β1Ar), and the human β-2 adrenergic (β2Ar) and A2A adenosine (A2Ar) receptors. We use our Protein Local Optimization Program (PLOP), which samples conformational space from first principles to build sets of loop candidates and then discriminates between them using our physics-based, all-atom energy function with implicit solvent. We also discuss a new kind of explicit membrane calculation developed for GPCR loops that interact, either in the native structure or in low-energy false-positive structures, with the membrane, and thus exist in a multiphase environment not previously incorporated in PLOP. Our results demonstrate a significant advance over previous work reported in the literature, and of particular note we are able to accurately restore the extremely long second extracellular loop (ECL2), which is also key for GPCR ligand binding. In the case of β2Ar, accurate ECL2 restoration required seeding a small helix into the loop in the appropriate region, based on alignment with the β1Ar ECL2 loop, and then running loop reconstruction simulations with and without the seeded helix present; simulations containing the helix attain significantly lower total energies than those without the helix, and have rmsds close to the native structure. For β1Ar, the same protocol was used, except the alignment was done to β2Ar. These results represent an encouraging start for the more difficult problem of accurate loop refinement for GPCR

homology modeling.

1.3 Loop prediction for a GPCR homology model: algorithms and results

Chapter 3 presents loop structure prediction results of the intra and extracellular loops of four G-protein coupled receptors (GPCRs): bovine rhodopsin (bRh), the turkey β1-adrenergic (β1Ar), the human β2-adrenergic (β2Ar) and the human A2a adenosine receptor (A2Ar) in perturbed environments. We used the Protein Local Optimization Program, which builds thousands of loop candidates by sampling rotamer states of the loops' constituent amino acid. The candidate loops are discriminated between with our physics-based, all-atom energy function, which is based on the OPLS force field with implicit solvent and also contains several correction terms. For relevant cases, explicit membrane molecules are included to simulate the effect of the membrane on loop structure. We also discuss a new sampling algorithm that divides phase space into different regions, allowing more thorough sampling of long loops that greatly improves results. In the first half of the paper, loop prediction is done with the GPCRs' transmembrane domains fixed in their crystallographic positions, while the loops are built one-by-one. Side chains near the loops are also in non-native conformations. The second half describes a full homology model of β2Ar using β1Ar as a template. No information about the crystal structure of β2Ar was used to build this homology model. We are able to capture the architecture of both short loops and the very long second extracellular loop, which is key for ligand binding. We believe this the first successful example of an RMSD validated, physics-based loop prediction in the context of a GPCR homology model.

1.4 Prediction of long loops with embedded secondary structure using the protein local optimization program

Chapter 4 closely examines the prediction of loops that contain small second structure

(loops or hairpins) segments, which is required for robust homology modeling at atomic-level accuracy. Particularly as loop prediction success extends to longer and longer loops, the exclusion of loops containing secondary structure becomes awkward. Here, we extend the applicability of the Protein Local Optimization Program (PLOP) to loops up to 17 residues in length that contain either helical or hairpin segments. In general, PLOP hierarchically samples conformational space and ranks candidate loops with a high-quality molecular mechanics force field. For loops identified to possess α-helical segments, we employ an alternative dihedral library composed of (ψ,ϕ) angles commonly found in helices. The alternative library is searched over a user-specified range of residues that define the helical bounds. The source of these helical bounds can be from popular secondary structure prediction software or from analysis of past loop predictions where a propensity to form a helix is observed. Due to the maturity of our energy model, the lowest energy loop across all experiments can be selected with an accuracy of sub-Ångström RMSD in 80% of cases, 1.0 to 1.5 Å RMSD in 14% of cases, and poorer than 1.5 Å RMSD in 6% of cases. The effectiveness of our current methods in predicting hairpin-containing loops is explored with hairpins up to 13 residues in length and again reaching an accuracy of sub-Ångström RMSD in 83% of cases, 1.0 to 1.5 Å RMSD in 10% of cases, and poorer than 1.5 Å RMSD in 7% of cases. Finally, we explore the effect of an imprecise surrounding environment, in which side chains, but not the backbone, are initially in perturbed geometries. In these cases, loops perturbed to 3Å RMSD from the native environment were restored to their native conformation with sub-Ångström RMSD.

1.5 Docking into the Kappa Opioid Receptor

Chapter 5 presents the results of docking actives and decoys in the kappa opioid receptor, and correctly separating them energetically. It also contains an explanation of the

binding mode of morphinans to the receptor (which is seen across the other crystallized opioid receptors as well). The successful docking is due to the integration of two key terms into the new scoring function, WScore. We motivate these terms physically, based on the water structure within the active site. The water structure comes from a WaterMap calculation. It must be noted that while this project has reached a publishable point, we believe that the results can still be improved, and we are actively working to do so. The final results should be published soon.

1.6 Details on the Protein Local Optimization Program

Each of the chapters have their own associated methods section that give a broad overview of the relevant methodologies used in the work, but their primary focus is on the new aspects added. Although the thesis is comprehensive with them alone, in this chapter, many addition details about PLOP are given as a reference for the reader. It includes a sample input file for clarification of many terms.

Chapter 2

Successful prediction of the intra- and extracellular loops for four G-Protein coupled receptors

2.1 Introduction

G-protein-coupled receptors, or GPCRs, are the largest class of membrane receptors in eukaryotes, and they account for more than 2% of the total genes encoded by the human genome(35). They are characterized by seven transmembrane (TM) helices, N-, and C-terminal fragments. The TM helices are connected by alternating intra- and extracellular loop regions that are very flexible and important for a wide range of biological functions (see Figure 2.1). Examples include mediation of most cellular responses to hormones, neurotransmitters, and chemokines. They are also responsible for blood pressure regulation, taste, vision, and olfaction (17). GPCRs activate heterotrimeric G proteins via agonist binding, which catalyzes GDP–GTP exchange. This acts as a molecular switch that, when turned on, modulates downstream effector proteins. It is estimated that GPCRs represent up to 50% of current pharmaceutical targets, which makes them extremely attractive candidates for rational drug design. Unfortunately, the development of therapeutics via structure-based design approaches that selectively target GPCRs has been severely impeded by the difficulty of obtaining accurate crystal structures at atomic resolution (36). In fact, as of the time of this study, there were only 17 (37) published crystal structures of six unique GPCRs: bovine rhodopsin (38), squid rhodopsin (sRh) (39), bovine opsin, the ligand-free form of rhodopsin, (Ops) (40), turkey β1- adrenergic receptor (β1AR) (41), human β2-adrenergic receptor (β2AR) (42), and human A2A adenosine receptor (A2Ar) (43). More recently,

the crystal structures of the CXCR4 chemokine and D3 dopamine receptor were published to the Protein Data Base (PDB). Thus, computational tools have been developed as an alternative approach to studying these key receptors.

Homology modeling has been the preferred method to build a structural model for a target protein from its sequence and the known structure of a homologous protein. However, this is a very difficult task for GPCRs, particularly because of the lack of structural homology in loop regions between currently known GPCR structures. This means that being able to generate accurate loop structures from ab initio principles would be helpful to the field. The 2008 GPCR dock competition (44), which attempted to assess the general state of GPCR structure modeling and ligand docking community-wide, demonstrated this well: It was determined that TM homology modeling can be done quite successfully, but the predicted loop regions were mostly poor. Even the best predictions for the second extracellular loop (ECL2) of A2Ar had a Cα root-mean-square deviation (rmsd) of more than 7 Å (45). Furthermore, the best predictions were actually done with de novo approaches. This had a profound impact on the accuracy of ligand-binding mode predictions and by extension has serious implications in drug design. A high quality, 3D model of the target GPCR is needed for 3D in silico screening of bioactive molecules, and it is well known that GPCR extracellular loops (ECLs) play an important role in high molecular weight peptidic ligand binding (46, 47). It has also recently been shown that ECLs (particularly ECL2) interact with low molecular weight ligands (i.e., adenosines, lipids, or biogenic amines) (48). ECL2 has also proven to be essential for GPCR activation (48).

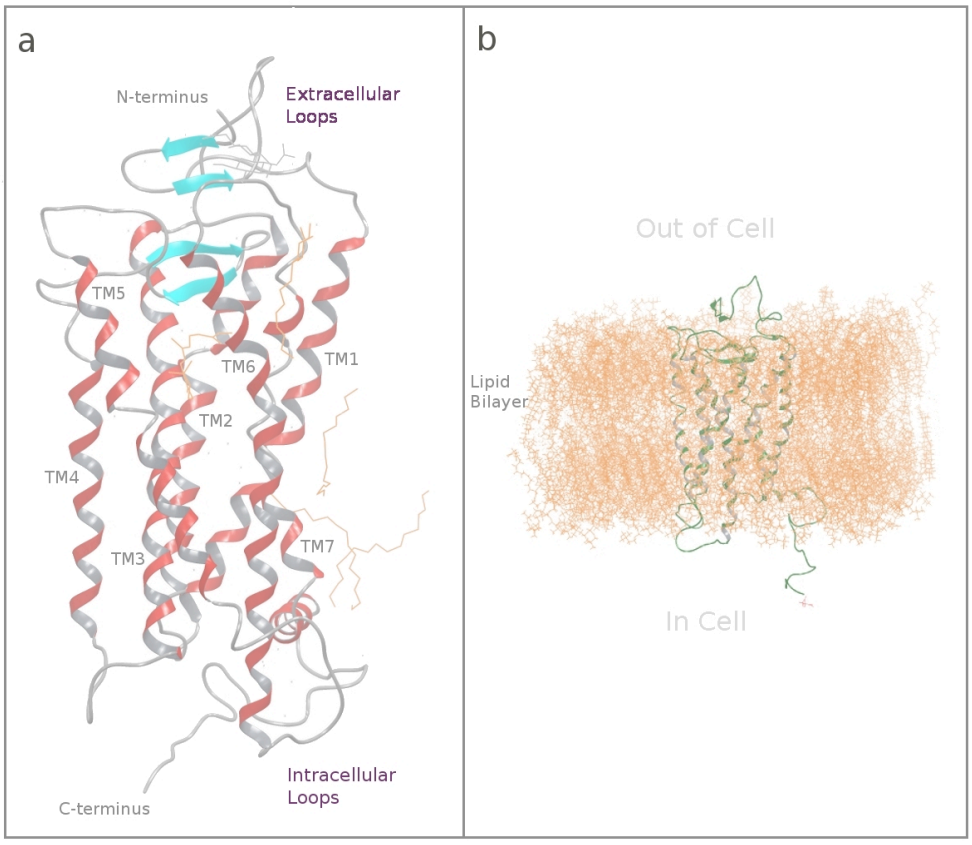
In addition to the importance of ECLs in ligand binding, intracellular loops

(ICLs) have been demonstrated to form key regions for G-protein coupling. Evidence suggests that the ICLs interact to form functional domains, which in turn interact with the G protein. They help to control receptor regulation through kinases, arrestins, and scaffolding proteins (49), and it is believed that the strength of interaction depends on ICL2 and the specificity on ICL3 (50). Perhaps even more striking are studies that show that when ICL2 and ICL3 are deleted, the GPCRs are no longer able to couple to G proteins while retaining their ligand binding conformations (51, 52). There are many other examples of point mutations to ICLs affecting the selectivity of GPCR binding to G proteins as reviewed in ref. (46).

It is clear that the ICLs and ECLs of GPCRs are of paramount importance to how they function, and that advances in modeling technology are needed to correctly predict their structure with computational methods. There has been extensive research on loop structure prediction over the past 20 y, and many programs for general loop prediction are available with a variety of features and accuracies (53-55). There has also been research directly focusing on loop modeling for GPCRs (46, 56, 57). Generally, long loops pose the greatest challenge, as conformational space increases exponentially with loop length, although even short loops can prove problematic. GPCRs present a further obstacle, in that many of their loops have significant interaction with the surrounding lipid bilayer. Whatever prediction method is being used needs to take into account the multiphase environment in which the loops are embedded. In this paper, we present loop restoration results of all of the ICLs and ECLs available for bRh (PDB ID code 1U19), β1AR (PDB ID code 2VT4), β2AR (PDB ID code 2RH1), and A2Ar (PDB ID code 3EML). We compare our results to prior studies in the literature (23, 24). We use an ab

initio methodology encoded in our Protein Local Optimization Program, otherwise known as PLOP (27, 58). Additionally, we deal with the multiphase properties of GPCRs for loops in which loop-membrane interactions significantly affect the loop structure with a unique approach described below. We are able to obtain excellent fidelity to the native loop structures for both short and (perhaps surprisingly) long loops, comparable to that obtained with our methods for soluble proteins.

Figure 2.1 Pictorial representations of GPCRs. a. A cartoon of bovine rhodopsin. The red sections are the 7 transmembrane helices (labeled TM1 through TM7), which are connected by gray, alternating intra and extracellular loops. The N- and C-termini are also labeled. b. A cartoon of bovine rhodopsin embedded within a lipid bilayer. When a ligand binds to the extracellular loops it induces a conformational shift of the receptor which triggers the intercellular domain to couple with a G-protein.



2.2 Results

We used PLOP to restore all of the crystallographically available loops of bRh, β1AR, β2Ar, and A2Ar, predicting each loop one at a time with the remaining loops fixed at either the crystallographic conformation or that obtained from a molecular dynamics (MD) simulation. The sequence, length, residue numbers, and rmsd of each loop are listed in Table 2.1. Eleven of the loops are considered short (5–7 residues), five medium (8–12 residues), and five superlong (over 15 residues). Thirteen out of the 21 predicted loops have an rmsd below 1 Å. This high precision is illustrated in Figure 2.2, which contains cartoons of the native (gray) and predicted (purple) ECL1s of each GPCR. As we can see, were it not for the different colors, they are practically indistinguishable from one another.

The restoration of long loops is a much more challenging endeavor that is necessary for working with GPCRs. The functional importance of the ECL2 in GPCRs has been demonstrated many times, and it is also consistently the longest loop. Table 2.1 displays the surprisingly high accuracy (given the difficulty of the problem) with which we were able to predict the structures of ECL2 for the four GPCRs in this study. In addition to their lengths ranging between 26 and 32 residues, the ECL2 from β1AR and β2AR contains a short helical fragment, and that from bRh possesses a region containing a β-hairpin structure. The crystal structure of ECL2 of A2Ar has missing residues (residues 149–155). We still predicted the structure of this loop, but because of the uncertainty of the native structure we consider this loop to be unsuitable for quantitatively calibrating accuracy. The native and predicted structures of the other three ECL2s are displayed in Figure 2.3. The restored structure of each of these

extremely long loops captures the folds and secondary structure fragments evident in the native structure. To facilitate getting the right secondary structure within the ECL2s, we employed a homology modeling-like approach, in which we identified that the center region of ECL2 could contain a helical portion. We then tested forcing a helix to form in that region versus a plain loop prediction and considered the structure with the lowest energy our final predicted loop. When forcing a helical region, PLOP samples a smaller set of backbone dihedral angles typical of α-helices for each residue in the helix. This is further elaborated upon in Methods.

Although most of the loop structures could be predicted with PLOP with less than 2-Å accuracy in our initial efforts employing the GPCR crystal structures and our standard continuum solvation protocol, some loops presented severe challenges; specifically, ICL2 of A2Ar and bRh, ECL2 of bRh, and ECL3 of A2Ar. Our hypothesis for these cases was that the loops in question interact significantly with the lipid bilayer, either the native conformation or in low energy, false-positive predictions, and the implicit solvent model in PLOP could not account for this multiphase environment. As an experiment, we built the explicit membrane for the two relevant proteins by running MD simulations and equilibrating the membranes with their respective receptors. We then reconstructed the loops in the presence of the lipids proximate to the loop. As we see in Table 2.1 the rmsds of loops predicted with the explicit membrane are significantly improved as compared to the corresponding calculation without any representation of the lipid bilayer. ICL3 of bRh is the one exceptional case: The crystal structure shows that although a small part of it interacts with the membrane, it is mostly stabilized by solvent. Thus, we did not believe that

imposing an explicit membrane would improve the predicted structure for two reasons. First, the majority of the loop is not lying in the membrane. Second, the MD loop and one of its flanking helices was largely divergent in conformation from the corresponding loop in the native structure, meaning that the predicted loop should not be exactly the same as the native. This is further buttressed by the fact that the rmsd of the native structure as compared to the MD structure of the loop is 8.8 Å. We nonetheless did the experiment, and our hypothesis proved correct. The rmsd of the predicted structure with the membrane as compared to the native was 8.80, almost the same as the prediction made without the membrane (8.51 Å). Additionally, the rmsd of the same predicted loop as compared to the MD loop was 4.01 Å, which, for this case, is a reasonable assessment of accuracy. All 18 residue loops are highly flexible, but this one poses a special complication in that its true structure is unclear given large discrepancy between the native and MD conformations. The status of this case (i.e., whether there is a serious problem with the energy model in PLOP, or whether the loop is extremely flexible and can occupy many diverse conformations in phase space with relatively low energetic penalty) will need to be further investigated in future work.

Figure 2.2 The first extracellular loop of the four representative GPCRs. In each case the native loop is gray, and the predicted loop is purple. The numbers denote the starting and ending points of each loop. a. The native and predicted structures of ECL1 of bRh; the backbone RMSD is 0.17Å. b. The native and predicted structures of ECL1 of A2Ar; the backbone RMSD is 0.18Å. c. The native and predicted structures of ECL1 of β1AR; the backbone RMSD is 0.27Å. d. The native and predicted structures of ECL1 of β2AR; the backbone RMSD is 0.12Å.

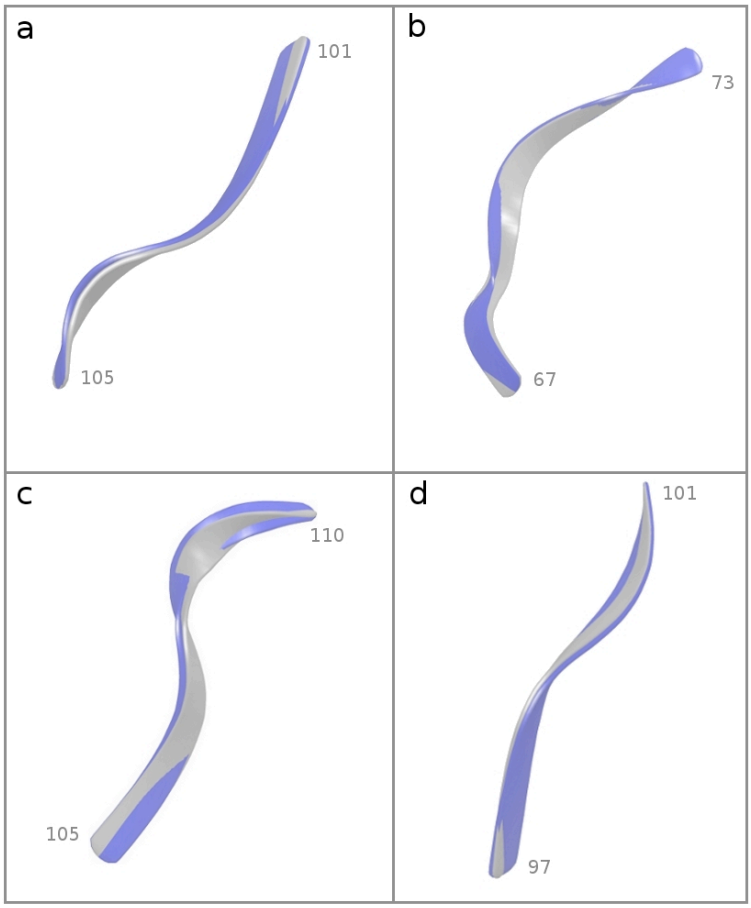


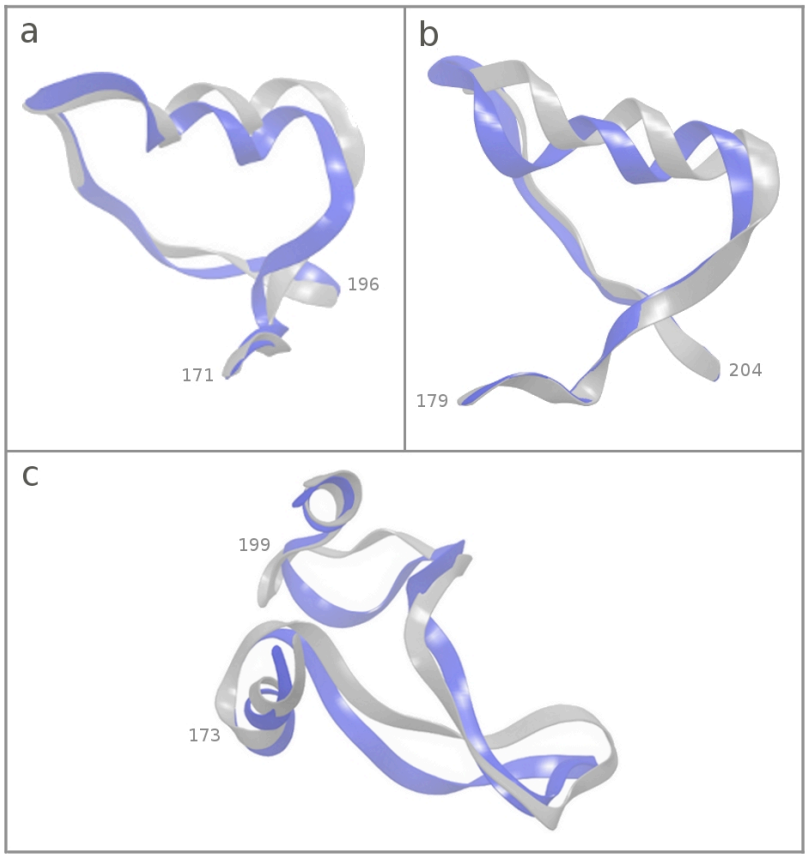
Table 2.1 The sequence of the six intracellular (ICL) and extracellular (ECL) loops of bovine rhodopsin, the human A2A adenosine receptor, turkey β1 adrenergic receptor and human β2 adrenergic receptor are listed here, except for ICL3 of A2Ar, β1AR and β2AR which, for crystallization purposes, is partially replaced by a T4 lysozyme. The RMSDs[a] are all of structures procured with methods already existing in PLOP and a set of parameters optimized for GPCRs. RMSD[b] Mem refers to the RMSD of the loop using the membrane method developed for this project. ECL2 of A2Ar (**) is missing 7 crystallographic residues. The RMSD is calculated using the residues specified by the crystal structure, while the missing residues are omitted in the calculation.

The RMSD and energy gap between intra and extracellular loops and their native counterparts

Loop	GPCR	Loop Sequence	Loop Length, Residue Numbering	RMSD[a] (Å)	RMSD[b] (Å) Mem
ECL1	**bRh**	GYFVF	**5**, (101-105)	**0.17**	
	A2Ar	STGFCAA	**7**, (67-73)	**0.18**	
	β1AR	GTWLWG	**6**, (105-110)	**0.27**	
	β2AR	KMWTF	**5**, (97-101)	**0.12**	
ECL2	**bRh**	VGWSRYIPEGMQCSCGIDYYTP HEETN	**27,** (173-199)	**11.53**	**3.44**
	A2Ar	GWNNCGQ(PKEGKNH)SQGCGE GQVACLFEDVVP	**32,** (142-173)**	**4.39**	
	β1AR	MHWWRDEDPQALKCYQDPGC CDFVTN	**26**, (179-204)	**1.59**	
	β2AR	MHWYRATHQEAINCYAEETCC DFFTN	**26,** (171-196)	**2.17**	
ECL3	**bRh**	HQGSDFG	**7**, (278-284)	**0.77**	
	A2Ar	CPDCSHAP	**8**, (259-266)	**1.94**	**1.11**
	β1AR	NRDLVP	**6**, (316-321)	**0.50**	
	β2AR	QDNLIR	**6**, (299-304)	**0.23**	
ICL1	**bRh**	HKKLRT	**6**, (65-70)	**0.41**	
	A2Ar	NSNLQNV	**7**, (34-40)	**0.35**	
	β1AR	TQRLQT	**6**, (69-74)	**0.78**	
	β2AR	FERLQT	**6**, (61-66)	**0.27**	

ICL2	**bRh**	CKPMSNFRFG	**10**, (140-149)	**5.79**	**2.86**
	A2Ar	RIPLRYNGLVT	**11**, (107-117)	**4.15**	**2.63**
	β1AR	ITSPFRYQSLMT	**12**, (143-154)	**0.33**	
	β2AR	SPFKYQSLLT	**10**. (137-146)	**0.46**	
ICL3	**bRh**	GQLVFTVKEAAAQQQESA	**18**, (224-241)	**8.51**	**8.80**
	A2Ar	Insertion of T4 lysozyme			
	β1AR	Insertion of T4 lysozyme			
	β2AR	Insertion of T4 lysozyme			

Figure 2.3 The ECL2s of β2AR, β1AR, and bRh. The native structures are gray, the predicted structures are purple, and the numbers denote the starting and ending residues of the respective loops. (A) The native and predicted structures of ECL2 of β2AR; the backbone rmsd is 2.17 Å. (B) The native and predicted structures of ECL2 of β1AR; the backbone rmsd is 1.59 Å. (C) The native and predicted structures of ECL2 of bRh; the backbone rmsd is 3.44 Å. This loop was predicted starting with an MD structure of bRh with explicit membrane molecules. It is compared to an aligned native structure. The flanking residues of the TM helices displayed in the cartoon superimpose very well, making this a meaningful comparison.



2.3 Discussion

In the past, PLOP has been tested on highly filtered sets of crystallographic loops in which the loop atoms had low average temperature *B* factors and real space *R* factors, very high resolution, and were far from a ligand. They also contained no secondary structure. These criteria ensured that efforts were focused on the development of a successful energy function and sampling strategy and not distracted by an imperfect crystal structure or interactions not described by the protein force field. Unfortunately, because of the difficulties in crystallizing membrane proteins, all of the GPCR loops modeled in the present paper violated one or more of these criteria. Furthermore, PLOP has never been used for membrane proteins, and it does not at present contain an extensively validated membrane model. Finally, several of the loops studied in the present paper are significantly longer than the loops on which PLOP has been extensively tested. These factors initially induced considerable uncertainty as to what sort of performance to expect with regard to accuracy for the set of GPCR loops studied here. However, as will be discussed below in detail, the results obtained provide quite good fidelity to the native structure in the great majority of cases, with precision and robustness comparable to what we have seen in our previous studies of soluble proteins.

PLOP uses a refined sampling grid, an all-atom physics-based energy function, and a careful side-chain packing algorithm that allows it to find and then pick out loops close to the native structure. However, it has previously only been optimized to deal with globular proteins, in which loops interact with aqueous solvent and other parts of the protein. We found that for the four GPCRs we studied, most of the loops are either very short or appear to be sitting on top of

the protein, primarily exposed to solvent and protein atoms as opposed to the lipid bilayer. For these cases, using PLOP with our previously optimized set of parameters (no parameters of the model, either in the force field or the continuum solvation component, were adjusted to improve the results of the calculations) was sufficient to produce excellent results. However, for cases in which the loop and membrane have important interactions, this was not sufficient. We postulated that the main source of error was the presence of a membrane interacting with a loop: A loop lying near the membrane has side chains poking into it, which gives that conformation favorable energetics. If, as in the calculation, solvent were to replace the membrane molecules, this conformation would no longer be energetically favorable. Thus, when running the prediction with the protein and the solvent, this conformation becomes a false negative. It cannot physically be the lowest energy structure when there is no membrane. The only way to find the correct structure was to in some way include the lipid bilayer into the calculation. Our solution to this problem involves using explicit membrane calculations (EMCs) in which three key torsional bonds of the lipid heads of membrane molecules within 7.5 Å of the target loop are sampled simultaneously as the loop is built up; this is described in depth in Methods. ICL2 and ECL3 of A2Ar both follow this hypothesis (see Figure 2.4*A*, which depicts A2Ar's ECL3 vs. ECL1 in membrane and solution, respectively). The native structure interacted with and was stabilized by the membrane. This is further buttressed by the fact that 25% of the contact points between ICL2 and the rest of the protein and membrane are with the membrane; even more strikingly, 55% of the contact points between ECL3 and all other possible atoms are with the lipid bilayer. The explicit membrane molecules also prevented the loop backbone from interpenetrating into the membrane region, a phenomenon seen when the membrane model was not present in the native structure. To gauge the potential bias the membrane molecules had on the loop prediction, we

looked at how much the lipid heads moved for the four loop reconstruction calculations for which the method was used. There were membrane molecules within 7.5 Å of ECL2 and ICL2 of bRh and of ICL2 of A2Ar, and their lipid heads were sampled. The rmsd between the starting conformation and the end conformation averaged over all of the mobile lipids was 1.90, 2.50, and 1.71 Å, respectively. The maximum rmsd for a single lipid head (of those sampled) between starting and end conformations were 3.25, 5.85, and 3.26 Å, respectively. Clearly, the lipid heads move significantly and do not greatly restrict the conformational freedom of the target loops. However, none of the membrane molecules were within 7.5 Å of the native ECL3 of A2Ar, meaning that none of the lipid heads were sampled while the loop was reconstructed. Despite the distance, the interaction energy between the membrane and the loop atoms is important, as the resultant loop has several interactions with the membrane, and, in this case, the immobile lipid heads may have biased the prediction more, because the membrane was optimized to the crystal structure as discussed before.

We also encountered two cases (ICL2 and ECL2 of bRh) where the native loop had little material contact with the membrane (zero contact points for ICL2 and only 1 out of 20 for ECL2), but the predicted structure without the membrane was occupying the membrane's space. Without explicit membrane molecules, the loop was found to be more energetically favored in this region than in its true position: A highly crowded protein environment rife with possibilities for steric clash. In this way, the absence of the membrane in the calculation produced a false positive due to the faux stabilization of solvent. ECL2 of bRh is a very long and folded loop, and it can thus easily extend into the membrane region if the membrane itself is not present. Additionally, 18 of the 27 amino acids that constitute this loop are polar, further supporting the idea that when

predicted only with protein and solvent, the solvent will provide a more attractive environment for the loop. When an explicit membrane was invoked, there was still enough space that the predicted structure could have avoided the crowded interior of the protein. Instead, PLOP correctly built and discriminated a final structure that is inside of the protein, with a good rmsd to experiment considering the length and complexity of the loop. To complicate matters more, as seen in Figure 2.4*B*, a small part of ECL2 of bRh is also near the membrane, making this case even more difficult and EMCs even more necessary.

It should be noted that for all four of these loop prediction calculations without the membrane, candidates closer to the native structures were found but were not lowest in energy. This further bolsters the idea that a low-energy, native-like structure cannot be found without an explicit membrane, when the structure either depends on the membrane or would occupy its space if it were replaced by solvent when crystallized. As discussed before, ICL3 of bRh does not fit either of these problematic states, and instead sits on top of the protein, mostly exposed to solvent; only a small portion has significant interaction with the membrane. Unsurprisingly, adding an explicit membrane to the calculation is unhelpful. Naturally, given only a sequence we will not generally know a priori where the loops of an unknown structure lie relative to the membrane and protein. Thus, as a precaution, we could use EMCs to predict the structure of each loop. It will only change the final predicted loop in cases that fit one of the scenarios described above.

Our results in comparison with work in the literature to date (56, 57) are shown in Table 2.2. We do not contrast our results with attempts at loop restoration made during homology modeling, as this is not a fair comparison. The exact coordinates of the flanking helices are

extremely important while building loops de novo and are not available with a homology model. Because we have an exact environment, our results would be biased positively. As described in ref. (56), Nikiforovich et al. use a de novo method with a coarse sampling grid to build candidate loops and do not incorporate water or the lipid membrane into their calculation. The results listed here are the rmsds that reflect the lowest energy structures from their calculation. Ultimately, when restoring loops for which we do not know the crystal structure, the only known way to choose a final answer from a bundle of loops is by choosing the lowest energy structure. For the short loops, our rmsds are considerably lower, and for the long loops, the differences are even more substantial. The results of Mehler et al. elucidated in detail in ref. (57) are comparable to ours in rmsd. The method they employ appears to be a promising one, but they do not present results for loops longer than seven residues, and the computational effort required even for these relatively straightforward cases, and how effort scales with loop length, are not discussed in ref. (57). The loop restoration calculations in this study ranged between 1.5 h for the shortest loops and 145 d for the longest loop with EMCs of single CPU time.

Figure 2.4 A2Ar and bRh with equilibrated membrane. (*A*) ECLs of A2Ar in membrane (gray molecules). The green loop is ECL3 and is buried in the membrane. The pink loop denotes ECL1 and is exposed to solvent. (*B*) ECLs of bRh in membrane (gray molecules). ECL2 is highlighted in pink, and it spans the protein, solvent, and lipid bilayer. A large portion of the loop is situated inside of the protein, a very crowded environment.

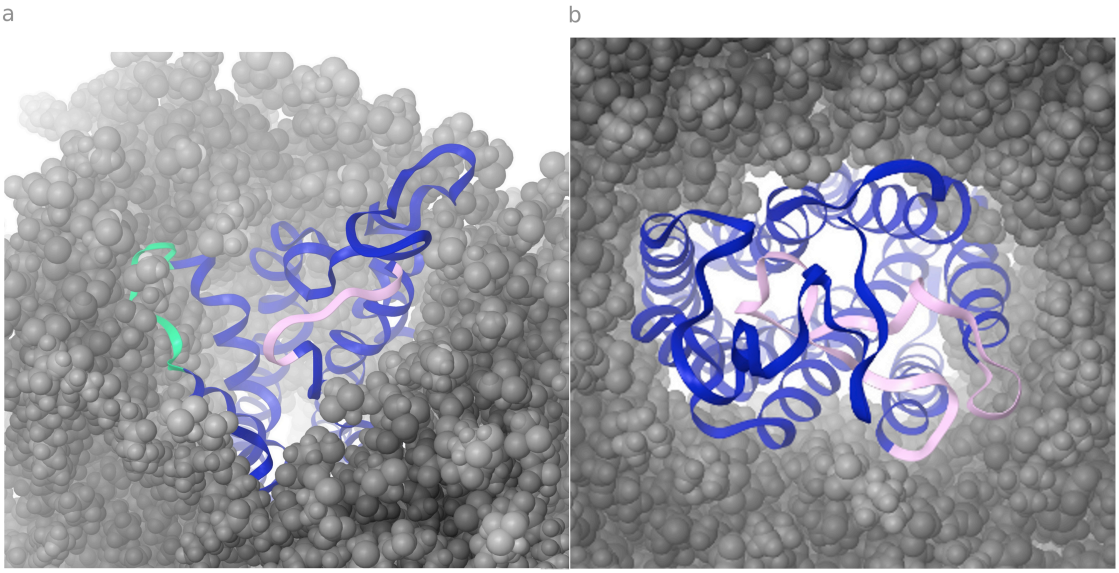


Table 2.2 Comparison of our results to those of similar studies. Ours are more accurate than those of Nikiforovich *et al* and comparable to those of Mehler *et al*. for the three short loops that they investigate. Note that the loop length of ECL2 of A2Ar is 32, where 7 of those residues do not have crystallographic data.

Comparison of our results to others in the literature					
Loop	**GPCR**	**Loop Length**	**RMSD[23] Å**	**C_a RMSD[24] Å**	**RMSD Å (this work)**
ECL1	bRh	5	5.2	0.37	0.17
	A2Ar	7	2.1		0.18
	β1AR	6	2.7		0.27
	β2AR	5	5.2		0.12
ECL2	bRh	27	7.4		3.44
	A2Ar	32(x7)	10.2		4.39
	β1AR	26	6.4		1.59
	β2AR	26	7.4		2.17
ECL3	bRh	7	2.8	0.55	0.77
	A2Ar	8	2.3		1.11
	β1AR	6	3.3		0.50
	β2AR	6	3.4		0.23
ICL1	bRh	6		0.44	0.41

2.4 Conclusions

Our goal was to restore the ICLs and ECLs of four GPCRs that were representative of the structures available while this study was being done. To do this, we utilized our loop-building program, PLOP, and developed a way to do EMCs to deal with cases in which a loop is spanning the solvent and membrane or is buried in the protein environment, but has an alternative (incorrect) low energy conformation occupying the space that should be taken up by the lipid bilayer. This combination yielded very good quality results for 20 out of the 21 loops for which crystallographic data exists. Our results represent a significant improvement as compared to what is currently reported in the published literature, and our procedure is able to handle loops ranging from very short to extremely long. Furthermore, because we only consider the lowest energy structure to be our final predicted loop, there will never be ambiguity as to which of the thousands of predicted structures to use for subsequent study. Thus, our method provides an excellent starting point for loop refinement in homology modeling.

Of course, the fact that we are able to predict loop structure one at a time, using the crystallographic coordinates (or coordinates generated by MD simulations starting from the crystal structure), is necessary, but not sufficient, evidence that we can successfully build loops in the context of a homology model. As the results of ref. (56) (which has a 2010 publication date) demonstrate, prediction of loops in GPCRs is very challenging even in the context of the native structure; it is also noteworthy that we were unable to find any efforts in the literature to restore GPCR loops in the context of the native structure longer than seven residues other than ref. (56). Although our results are highly encouraging, a demonstration of practically useful

prediction machinery will require starting from a homology model and achieving results of a similar quality. This is a substantially more challenging sampling problem than what we have undertaken here; on the other hand, our calculations utilize very little computer time by modern standards, and have a correspondingly low financial cost (given that a single processing core can be purchased for $250, the effective cost of predicting even a 26-residue ECL2 is approximately $60). For realistic GPCR homology model refinement, one can envision deploying many orders of magnitude more computer time, given the importance of the problem, utilizing more global algorithms in which our highly efficient localized prediction methods are embedded and play a critical role. Work along these lines is currently in progress in our laboratory. Ultimately, the development of a set of tools that can accurately and consistently be used for homology modeling is essential for future drug design work for GPCRs and many other protein families as well.

2.5 Methods

The computational techniques used in this paper for loop restoration have been described in great detail elsewhere (27, 58), but we provide a brief overview of the method here. We also describe an addition to the methodology that allowed us to deal with cases where the membrane plays a key role in determining loop structure. PLOP contains a single loop prediction algorithm in which a set of loop conformations are generated by an ab initio phase space search of possible loop geometries, screened for obviously poor interactions and then clustered and scored via an all-atom energy function with implicit solvent. Crystal neighbors are used in all calculations where the membrane is not included. The location of crystal contacts (atoms within 4 Å of one another) are found in Table 2.3. Conformational space is spanned via a dihedral angle

search that samples combinations of dihedral angles (ϕ,ψ) (a discretized Ramachandran plot) for each natural amino acid. Of course, it is too computationally expensive to sample every single backbone dihedral angle combination for a loop of nontrivial length; thus, quick screening techniques are used to attack this problem. First, the candidates are rejected if they fail the hard sphere steric clash check, which relies on an overlap factor (ofac). The ofac is defined as the ratio of the distance between two atoms to the sum of their van der Waals radii. Although the default ofac in PLOP is set to 0.70, for GPCRs it was found that a lower value of 0.55, which allows more loop candidates to be generated, was preferable. The remaining thousands of loop candidates are then clustered based on structural redundancy, and representative loops (closest to the cluster center) are chosen. This final set of loops is then optimized and scored using an energy function based on the Optimized Potential for Liquid Simulations all-atom force field and the Surface Generalized Born model of polar solvation. The energy function has been optimized for protein side chain and loop predictions with a variety of corrections such as a hydrophobic term adapted from the ChemScore scoring function, and the variable dielectric model (59). The lowest energy loop is the final predicted loop of this single loop prediction.

For long loops (13 or more residues) the same general scheme for a single loop calculation is followed, but there are some major differences that make it computationally viable. The biggest change lies in the dihedral angle sampling. For short loops, the (ϕ,ψ) angles for each residue are sampled, but for long loops, dipeptide sampling is used based on a library of sets of five consecutive dihedral angles (ϕ_1,ψ_1,ω,ϕ_2,ψ_2). This effectively reduces the number of possible combinations of residue positioning and allows us to search loop space in a way that is

computationally accessible.

A full loop prediction involves a hierarchy of stages, each of which contains multiple single loop predictions. For short loops, in the initial stage (Init), five single loop predictions are done with five different ofacs (0.45, 0.50, 0.55, 0.60, and 0.65). The top five loop candidates from each of these loop calculations are then passed on to the first refinement (Ref1) stage, in which each model is subjected to further sampling using a Cartesian constraint of 4 Å on each Cα atom. This allows us to do finer sampling around these energy minima. The loops with the lowest energies from both the Init and Ref1 stage are then passed onto the refinement 2 (Ref2) stage, where they are constrained by 2 Å on each Cα atom. Finally, the loop that has the lowest energy from all stages is the predicted loop structure, and its rmsd is calculated using the N, Cα, and C atoms in the loop backbone. We report global rmsds, meaning that the body of the predicted structure is superimposed on the body of the native structure (as opposed to superimposing loops locally), and then the rmsd of the loop atoms is calculated. The same hierarchical approach to loop prediction is applied to long loops (greater than 13 residues); however, between the Ref1 and Ref2 stages there are a series of fixed stages in which beginning and ending residues are fixed in space (the number of fixed residues increases with each fixed stage), and the remaining center fragment of the target loop is sampled. As before, after each fixed stage, the lowest energy loops from all stages up to the current one are passed onto the next stage. It was found that for GPCRs six fixed stages was sufficient.

For loops that contain helical fragments, we use a modified version of PLOP, in which the helix residues are treated as one special residue that is sampled and built up like a normal amino acid. Once loop candidates are generated, a helix is formed in this special region of the loop based on a separate library of helix backbone dihedral angles, thus imposing a helix in the

loop. A manuscript presenting a more complete treatment of this methodology, with a large number of test cases taken from soluble proteins in the PDB, is currently in preparation. Lastly, for loops that are poking into the membrane and whose conformations are utterly inseparable from membrane-loop interactions, we employ a special procedure in which the explicit membrane was included in the calculation, which we term EMCs. The membrane structures and placement for bovine rhodopsin came from a 250-ns all-atom explicit solvent simulation run with CHARMM and was done by George Khelashvili and Harel Weinstein (60). The human A2A adenosine receptor was similarly run by Schrodinger, Inc. for 930 ns with AMBER and the amber99 force field (61). We aligned the MD protein structure with the native, and as the key regions near the target loops were very similar, we then ran the loop prediction on the MD structure. Additionally, up to three key torsional bonds of the rotating lipid heads of the membrane molecules were sampled together with all side chains within 7.5 Å of the loop (62). The goal was to capture the fluid properties of a membrane as well as bias the prediction as little as possible. We allow the side chains of the loops to fit into the membrane, which is also being sampled simultaneously as opposed to blocked entirely from entrance into certain spots of the membrane. The MD structure of the protein is then superimposed on the crystal structure, and the global backbone rmsd between the predicted and native loop is found. Although this is not a perfect comparison, the flanking helices overlap sufficiently well that the rmsd calculation is certainly meaningful. In this paper, the membrane region was optimized to the full native protein that significantly relaxed and moved throughout the simulation. Thus, the position of membrane near loops is less biased than if the protein had been held still during the MD simulation. Furthermore, any “correct loop inducing” effect would not have affected most of the loops because they are immersed in solution, far away from the membrane molecules. In a future

publication, we will address GPCR loop prediction in a membrane environment that was only optimized to the TM regions.

The method used to predict loops containing helical fragments requires an initial guess for the position of the helical fragment; it also necessitates comparison of the helical structure with possible alternatives in which a helix is not formed. The latter is readily accomplished by running two simulations, one normal simulation that does not specify a helical library for a particular region, and a second that does, and comparing the total energies of the two simulations to select the final prediction. The former issue is complicated in the general case. One approach is to use one or more secondary structure prediction methods to predict the position of a putative helical region. This approach succeeds in many cases, as we will describe in a subsequent publication. For the present specific case under study (the ECL2 loop in GPCRs), secondary structure prediction from PSIPRED (63) does not yield a helical fragment for any of the GPCRs we investigated. However, given a database of known GPCR structures, and the objective of building homology models of the remaining structures, a straightforward alternative is to align the target ECL2 loop with the other known structures and try a helical fragment in the region derived from the alignment (assuming a helix exists in at least one of the other loops). For example, we used this approach to test the validity of the prediction for the ECL2 loop of bRh, which does not contain a helical fragment. The ECL2 loop of bRh was aligned to the same loop in β1AR and β2AR, and a helical fragment library built in at the indicated position in the bRh/ECL2 loop prediction. The result of this calculation yielded an energy that was significantly higher (by 38.2 kcal/mol) than the normal calculation, thus yielding the correct structure for this system. Similarly, the simulations containing helical libraries for the remaining two cases correctly yielded lower energies (by 18.5 kcal/mol at a minimum) as compared to normal

simulations. Thus, although this approach needs to be tested for realistic homology modeling cases to be fully validated, the initial results satisfy all of the relevant success criteria, and there are no obvious reasons why similar success cannot be achieved for more challenging problems (in some cases, multiple simulations in which, for example, the helix fragment length is varied may be required, but this necessitates an acceptable small integer multiple increase in computation time).

Table 2.3 All loop predictions that did not include an explicit membrane were performed using crystal neighbors in the calculations. All the copies of the asymmetric unit are predicted simultaneously, rather than the entire structures of the neighbors being used to guide the prediction of the central asymmetric unit. Listed above are the crystal contacts (defined as being within 4Å) that exist in the 21 loops that we predicted for this paper.

Crystal Contacts				
GPCR	**Loop**	**Residue 1 Name_Number: Atom Name**	**Residue 2 Name_Number: Atom Type**	**Distance (Å)**
β2AR	171-196	ASN_1053: OD1	ILE_177: O	2.91
		THR_1109: CG2	GLU_187: OE1	3.56
	299-304	LYS_1035: O	ASN_301: CB	3.14
		PRO_1037: CG	GLN_299: NE2	3.4
bRh	224-241	THR_242: OG1	GLN_237: NE2	2.95
		GLN_237: O	GLN_238: CB	2.83
		GLN_236: OE1	GLN_238: OE1	2.33
		GLN_238: O	GLN_238: OE1	2.92
	173-199	GLU_196: OE2	PRO_194: CG	3.02
A2Ar	107-117	LYS_292: CE	ASN_113: ND2	3.34
	142-173	GLN_207: NE2	GLN_157: O	2.83
		ARG_369: NH2	GLY_158: O	3.05
		ASN_247: OD1	GLU_161: N	3.26
		LYS_250: CB	GLU_161: OE1	3.07
		SER_251: OG	GLU_161: CB	3.73
		ASN_260: OD1	GLY_147: OE2	3.13
		ASN_262: OD1	GLY_162: CA	3.05

	259-266	GLU_161: OE2	ASN_260: OD1	3.13
	34-40	THR_316: OG1	SER_35: CB	3.41
		GLU_315: CB	SER_35: OG	3.64
		ASN_288: ND2	GLN_38: OE1	2.75
	67-73	LEU_202: CD2	PHE_70: CG	3.54
		ARG_205: O	CYS_71: N	2.72

Chapter 3

Loop prediction for a GPCR homology model: algorithms and results

3.1 Introduction

Computational modeling of three-dimensional (3D) protein structures can facilitate structure-based drug design, the modeling of protein-ligand and protein-protein interactions, or, perhaps someday even rational design of novel proteins (64-67). G-protein coupled receptors, or GPCRs, mediate responses to a vast number and variety of bioactive molecules (1). They represent an exceptionally important class of receptors to be modeled, as they are crucial to basic physiological functions ranging from neurotransmission to cell growth to blood pressure regulation(21). They are implicated in many human pathologies such as tumor metastasis and Alzheimer's disease(22). They are already the targets of more than 50% of pharmaceutical compounds, making them one of, if not the most, valuable classes of drug targets in the human body.

Computational studies are valuable for probing GPCR ligand binding as well as structure and function questions. They rely on an x-ray crystal structure, which can be difficult, if not impossible, to obtain for a GPCR of interest (68). An alternative approach to determining 3D protein structures is homology modeling (HM) (69). The basic architecture of building a homology model is as follows. If we know the structure arising from one sequence of amino acids (sequence A, our template) and have another similar sequence (B, the target), we can use our knowledge of the structure of A to predict the structure of B. To do this one aligns the two sequences based on sequence identity, supplemented by specific points of alignment for key conserved residues, generates the backbone regions of the target based on these residues' positions in the template, and lastly models the flexible loop regions as well as side chains. The

theoretical basis for this approach lies in the fact that protein structure is uniquely determined by amino acid sequence, and that, throughout evolution, protein sequences changed at a much faster rate stable structures (70). Thus, highly similar sequences fold into practically identical structures while even structures with low but significant (~20-30%) sequence identity will still typically fold into related structures.

GPCRs all have a common architecture of seven transmembrane (TM) helices connected by alternating intracellular and extracellular loops. For sufficiently high sequence identity, the TM helix region will ordinarily yield very good alignments, with relatively small deviations in the structure of the target as compared to the template, thus making it straightforward to build good (although not perfect) models of this region of the target using standard homology modeling methods. However, the loop regions can vary considerably, particularly in the 30-40% sequence identity range (ie. between GPCR sub-classes), where much of the interesting and relevant homology for GPCRs is to be found. Hence, refinement of the loop regions is potentially critical in constructing accurate homology models for many GPCRs whose structures have not yet been determined experimentally.

The Protein Local Optimization Program (PLOP) has been developed over the last decade to refine flexible regions of globular proteins (27, 28, 58, 71). Until recently, most efforts were made to accurately reconstruct loop structure in the native crystal structure environment. In our most comprehensive test of PLOP, the restoration of 115 loops between 14 and 20 residues resulted in an average backbone root mean squared deviation (RMSD) of 0.82Å (28). We also demonstrated that we were able to restore the intra and extracellular loops of four GPCRs in their native environments: the bovine rhodopsin (bRh), turkey β1-adrenergic (β1Ar), human β2-adrenergic (β2Ar) and human A2a adenosine receptor (A2Ar) (30). There are other programs

that aim to accurately predict loop structure. A few examples are: Rosetta (55), which, like PLOP, is an *ab initio* method that constructs loops from backbone dihedral angles and short peptide fragments and ranks the structures with a physical chemistry based scoring function; SuperLooper (72), which is a knowledge based method and does the predictions by slelecting loops from a large database of known loop structures; and Modeller,(53) which combines *ab initio loop construction* and database extracted knowledge. Although each of these programs has strengths and weaknesses, none have been extensively tested in loop prediction of GPCRs. Successful results in the context of native structures are a prerequisite for being able to refine in a homology model, and while the results were encouraging, we did not know to what extent modifications to the rest of the protein structure would affect loop prediction accuracy.

To investigate the effects of small structural perturbations in a loop's environment we opted to continue our studies of the GPCRs we had already worked with extensively. As a first test, we sought to restore the same set of loops, this time, however, only leaving the TM bundle residues in their native positions, with the rest of the loops deleted. These results constitute the first half of this paper. The second half focuses on predictions in the context of a homology model of the human β2 adrenergic receptor built from the turkey β1 adrenergic receptor template. These two receptors have 62% sequence identity, as opposed to as low as 15% between other known GPCR pairs, and highly similar structures, ensuring that the perturbations in the backbone of the TM bundle are not too large. The loops of the two receptors are also quite similar in structure, and thus the homology model target loops based on template loops are close to the experimentally determined loops of the target crystal structure. However, even in a case like this, *ab initio* loop prediction on the same homology model is far from guaranteed to produce accurate loops. The homology model potentially introduces a slew of inaccuracies in

nearby side chains, errors in the loop stems' backbone, and errors in the backbone/side chains in the rest of the protein. Loop restoration in the exact crystal environment can be thought of as solving a localized jigsaw puzzle, in which the crystal structure solution and myriad of extremely similar solutions have to fit the constraints of the precise environment. As soon as that environment is changed, it is possible that there are alternative low energy structures that constitute local minima in the particular energy model that is being used to score the various protein conformations. Therefore, the problem essentially becomes a larger jigsaw puzzle, in which more than just the local loop region has to be extensively sampled.

We show below that, with increased sampling, we are able to achieve accurate loop refinement of the homology model. The results reflects the robustness of PLOP and benefit from a high degree of similarity in the TM domain. Success in *ab initio* reconstruction of the loops in a homology model derived from a template with a sequence identity of 62%, where the results without any refinement have low RMSDs to the native structure, does not prove unambiguously that such reconstruction would be equally successful in a more challenging case, for example one in which the sequence identity is only ~35% rather than 62%. However, in order to test both the energy model and sampling algorithms in our refinement methodology, it is necessary to proceed incrementally. The results shown here do represent the next milestone in the ability to apply PLOP based refinement algorithms to GPCR homology modeling efforts of practical interest. Indeed, our next project will directly address a target/template pair in the lower (but still tractable) sequence homology range.

3.2 Results and Discussion

3.2.1 Loop prediction in an imperfect environment

For the first part of this project, we used PLOP to predict the extracellular and

intracellular loops (ECLs and ICLs, respectively) of bRh (PDB ID code 1U19) (38), β1Ar (PDB ID code 2VT4) (41), β2Ar (PDB ID code 2RH1) (42), and A2Ar (PDB ID code 3EML) (43). The TM bundle residues are fixed in the crystallographic conformation, or at the conformation obtained from the explicit-membrane molecular dynamics (MD) simulation, in which all protein non-hydrogen atoms were tightly constrained. Consequently, the location of the TM bundle residues are almost identical in both loop prediction calculations run with and without a membrane present. The flexible regions of the proteins—the loops and the N- and C-terminal tails—were removed. The T4 lysozyme which takes the place of ICL3 in A2Ar, β1Ar, and β2Ar was also removed. The loops were then predicted, one-by-one, on both the extracellular and intracellular domains of the protein. To further perturb the local environment, side chains within 7.5Å of the target loop were predicted simultaneously. First, the shortest loop of the extracellular domain (typically ECL1) was reconstructed. Then the middle-length loop (typically ECL3) was reconstructed with the predicted structure of ECL1 kept in place. Finally, ECL2 was reconstructed with the predicted positions of ECL1 and ECL3 in place. The same general scheme was used for the intracellular domain, only in this case, because we did not have crystal coordinates of ICL3 for most of the proteins, the short ICL1 was first reconstructed, and then ICL2 was predicted given ICL1's predicted position. In our previous work we had unresolved trouble predicting the structure of ICL3 of bRh, even in the native crystal structure. We are still trying to understand this case, but for this study, we wanted only to proceed with these more challenging calculations by using loops that we knew could be accurately reconstructed in the native structure. Table 3.1 contains the sequence, length, residue numbers, and root mean squared deviation (RMSD) of each loop. Figure 2.1 panels a and b contain cartoon pictures of all of the intra and extracellular loops of β2Ar and bRh (situated in the membrane), which illustrates

the range of RMSDs cited in Table 3.1, ranging from 0.13Å to 6.29Å.

In previous work, we demonstrated that we could predict the structure of the intra and extracellular loops of these four GPCRs. The difference between that study and the present one is that in the previous investigation, the loops were predicted in an environment incorporating the crystallographic conformation of all of the other residues, including the loops, tails, and nearby side chains, whereas in the present case, we do not assume knowledge of any of the native loop conformations, or their surrounding side chains. When building a real homology model, the native locations of all of the residues are uncertain, and thus significant noise is introduced to the system. Our strategy is to build up to the full problem in stages; the environment described above is not as challenging as a realistic homology model environment would be, but it is significantly more challenging than a perfect native environment. As with the case of the predictions in the native environment, success in this endeavor, is necessary, but not sufficient, to attempt prediction in an actual homology environment. An advantage of this incremental approach is that the errors made in such an intermediate level of calculations can be more easily dissected than those in an environment where errors in loop prediction could be due to many different types of structural discrepancies.

By focusing first on columns 5 and 6 of Table 3.1, we see that all predicted loops of length 5-7 are in excellent agreement with the experimentally determined loop structures. These loops have an average RMDS of 0.34Å, essentially identical to the errors reported in our previous work where the same set of loops' average RMSD is 0.36Å. These loops were expected to be predicted with similar accuracy, because, in addition to the fact they are short, they are relatively extended as compared to the distance between loop stems, a type of loop structure that we have found to be easier to predict accurately than those in which the maximum loop length is

significantly larger than the distance between stems, a situation that allows more "play" in the loop. Furthermore, interactions with nearby loops, clearly contributes in only a limited fashion to the energies of these various predicted loops. This is not true for the longest loops, which do seem to depend on having reasonably good predictions of the short, nearby loops. The intermediate length loops, ICL2 of all four receptors and ECL3 of A2Ar, appear to also behave similarly to what we saw in our previous work, and we reserve further analysis of ICL2 of bRh and A2Ar and ECL3 of A2Ar for later in this paper.

The longest loops, ECL2 of all four receptors, proved to be more challenging in this study than in the previous one, although ECL2 of A2Ar is actually predicted to higher accuracy. Preliminary testing pointed to evidence that small changes in the environment could, for example, cause the prediction of ECL2 in β2Ar to go from 2.17Å in the native case to 6.10Å. We realized that we would need to more extensively sample phase space if PLOP were to find a conformation that would be close to the native loop and also lowest in energy. Once the environment is changed, due to errors in the other loop predictions, these small perturbations can make structures that are close to the native have artificially higher energies, due to steric clashes that would not form in the native. The same is true for short loops, but because their conformational flexibility is significantly more limited than very long loops, their sampling is already exhausted by our original methods. To alleviate the effects of these clashes on final loop selection, we will have to continue to work on new sampling algorithms. We also have a new term being parameterized that penalizes loops (and, in the future, their nearby environment), that contain dipeptide rotamers not commonly found in nature. For the purposes of this work, our new phase space partitioning algorithm described in the Materials and Methods section was able sufficiently sample phase space such that our predicted loop have similar fidelity to the native

loops as in the original GPCR loop study. Furthermore, as discussed in the Materials and Methods section, for the prediction of ECL2 of β1Ar and β2Ar, a homology modeling-like approach was taken to ensure that the small helix in both structures is formed. We run loop predictions with both the helical region (as determined by aligning the loops and using the known helical residues from one as a guess for the helical region for the other) enforced and without any constraints. When the loops are run without the constraint, the prediction for ECL2 of β1Ar is 5.62Å (as compared to 2.73Å), and is also 52.57 kcal higher in energy. The RMSD of ECL2 of β2Ar without a helical constraint is 13.76Å (as compared to 2.16) and 7.06 kcal higher in energy. Thus, the "best RMSD" structures also correspond to the lowest energy prediction in both cases.

ECL2 of A2Ar, as said before, improved despite the more difficult loop-building environment. This particular loop has seven missing residues in the crystal structure, and we only predict the residues for which we have crystallographic data. Given that the predicted ECL1 and ECL3 of A2Ar are in such close agreement with the crystal structure (and certainly within experimental error), the environment in which we predict ECL2 is extremely close to the same prediction in the native protein. We attribute the significantly better prediction (2.92Å as compared to 4.39Å RMSD) to the use of phase space partitioning screening. Improvements in the new VSGB2.0 energy model which were not available when running the loop predictions in our previous GPCR work may have also contributed positively to the prediction. ECL2 of bRh remains the hardest of the four ECL2s that we attempted, and to discuss it thoroughly we must first discuss explicit membrane calculations (EMCs) developed for GPCRs in our first paper.

In the original work, there were four cases, ECL2 of bRh, ECL3 of A2Ar, and ICL2 of bRh and A2Ar, which had errors if we did not explicitly include membrane molecules into the

simulation. In the cases of ICL2 and ECL2 of bRh, the predicted loops were occupying regions of space taken up by the membrane. They falsely gained energetic stability from the interactions with solvent, where in reality this area is occupied by membrane molecules. These loop positions should have instead incurred a high energy penalty from attempting to bury the loop in the lipid bilayer. Conversely, ECL3 and ICL2 of A2Ar interact with nearby lipid heads of the bilayer. Without explicit membrane molecules included in the calculation, it was impossible for the correct conformations of the loops to gain favorable energetics from the loop residue-membrane molecule interactions.

We assumed that for these four loops, EMCs would still be required. As seen in Table 3.1, inclusion of explicit membrane molecules significantly improves predicted loop results in the imperfect environment. However, we do see for ICL2 of bRh and A2ar a small degradation in results as compared to predictions done in the fully native environment. This is most likely due the added complexity of the non-native ICL1 and nearby side chain conformations as well as the fact that the membrane in this case is not equilibrated in the presence of the native loop. This invokes an important unanswered question from the previous study involving whether or not inclusion of explicit membrane molecules could cause predictions of well solvated loops to become worse. To determine this, we predicted all the loops of bRh and A2Ar both with and without an explicit membrane as a way to calibrate its effects on loops that are restored well without it. The one exception is ECL2 of A2Ar; since there are missing residues in this loop in the crystal structure, we do not consider this a good loop to calibrate accuracy of our methods, particularly when adding the computational complexity of nearby rotating lipid heads. The EMCs for this part of the study, as described in more detail in the Materials and Methods section, differ from our previous work with GPCRs in that the membrane is equilibrated only to the

native TM bundle. Previously, it was equilibrated with the native loop positions in place, although the key torsional bonds of the membrane molecules' lipid heads that have significant interactions with the loops are rotated (and move significantly) in an attempt to prevent the conformational freedom of the target loops. In the current study, this potentially positive bias in loop prediction is eliminated.

Instead, a new potential problem is introduced: because the membrane is equilibrated around only the TM helices, it may inhibit correct loop conformation, which appears to occur with ECL1 of A2Ar. In the superimposed native structure, the lipophilic side chains, particularly residue Phe70, are poking down into the membrane in physically impossible positions. For example, several carbons on Phe70 are around 1Å away from membrane carbon atoms. When the loop is predicted using an EMC, it positions itself such that these carbons can have favorable interactions with the membrane (ie. around 3.5Å). The same problem is affecting residue Thr68. If the membrane were not equilibrated so close to the loop area, then this loop should have a final predicted structure that is close to the calculation done without the explicit membrane and agrees well with the crystal structure. Unfortunately, this is an unavoidable issue: if one does not have a good guess for loop structures, the best way to equilibrate the membrane is around the TM domain. Thus, the total number of degrees of freedom is much higher, and even when sampling the lipid heads (thereby giving them freedom to allow some reasonable loop to form), we expect to see some degradation in results as compared to our previous work. Nonetheless, even in the case of ECL1 of A2Ar, the final prediction is still quite reasonable (see Figure 3.2).

ECL2 of bRh is the only loop for which we were not able to obtain a comparable predicted loop as in the original work. The 9.14Å RMSD cited in Table 3.1 represents the result using the same phase space partitioning that we found useful in the other cases (screening for the

loop closure residue residing in one of four quadrants). We attempted to add in the membrane in two ways. The first was with a modified phase space partitioning, in which we attempted to put a plane tangent to where the membrane comes up across the endpoints of the loop. This resulted in a loop with a 7.82Å RMSD. Inclusion of a full explicit membrane improved the prediction to an RMSD of 6.29Å, reaffirming that including the lipid molecules into the electrostatics is important. However, we are still currently unable to obtain an accurate loop for this prediction. ECL2 of the adrenergic and adenosine receptors are well solvated and sticking up on top of the protein. ECL2 of bRh is contained entirely within and interacts heavily with the extracellular domain of the protein and is thus going to be even more sensitive to changes nearby. This loop will serve as an excellent test case for future research, as improving its prediction will signify an important step forward for the homology modeling methodology.

Nevertheless, overall, we are able to obtain predicted loops with excellent fidelity to their corresponding native loop structures, despite the imperfect environment.

Table 3.1 The sequence and numbering of the ICL and ECL loops of bovine rhodopsin, the human A2A adenosine receptor, turkey β1 and human β2 adrenergic receptor are listed, along with the corresponding RMSDs of predicted loops compared to their native counterparts. RMSD[a] refers to plain loop prediction, while the values in the RMSD[b] column are garnered by explicit membrane calculations. Residues 8-14 of ECL2 of A2Ar are missing in the crystal structure, the RMSD is calculated only using the known atomic coordinates. The RMSDs of ECL2 of β1Ar and β2Ar correspond to our lowest energy prediction, and are accomplished by means of a helical constraint enforced during loop prediction.

The RMSD between intra and extracellular loops and their native counterparts					
Loop	**GPCR**	**Loop Sequence**	**Loop Length, Residue Numbering**	**RMSD[a] (Å)**	**RMSD [b] (Å)**
ECL1	**bRh**	GYFVF	**5**, (101-105)	**0.15**	**0.26**
	A2Ar	STGFCAA	**7**, (67-73)	**0.26**	**1.78**
	β1AR	GTWLWG	**6**, (105-110)	**0.20**	
	β2AR	KMWTF	**5**, (97-101)	**0.13**	
ECL2	**bRh**	VGWSRYIPEGMQCSCGIDYYTPHEETN	**27,** (173-199)	**9.14**	**6.29**
	A2Ar	GWNNCGQ(PKEGKNH)SQGCGEGQVACLFEDVVP	**32,** (142-173)**	**2.92**	
	β1AR	MHWWRDEDPQALKCYQDPGCCDFVTN	**26**, (179-204)	**2.73**	
	β2AR	MHWYRATHQEAINCYAEETCCDFFTN	**26,** (171-196)	**2.16**	
ECL3	**bRh**	HQGSDFG	**7**, (278-284)	**0.48**	**0.52**
	A2Ar	CPDCSHAP	**8**, (259-266)	**1.90**	**1.47**
	β1AR	NRDLVP	**6**, (316-321)	**0.68**	
	β2AR	QDNLIR	**6**, (299-304)	**0.25**	
ICL1	**bRh**	HKKLRT	**6**, (65-70)	**0.32**	**0.43**
	A2Ar	NSNLQNV	**7**, (34-40)	**0.40**	**0.33**
	β1AR	TQRLQT	**6**, (69-74)	**0.47**	
	β2AR	FERLQT	**6**, (61-66)	**0.36**	
ICL2	**bRh**	CKPMSNFRFG	**10**, (140-149)	**6.90**	**3.91**
	A2Ar	RIPLRYNGLVT	**11**, (107-117)	**3.74**	**2.73**
	β1AR	ITSPFRYQSLMT	**12**, (143-154)	**1.27**	
	β2AR	SPFKYQSLLT	**10**. (137-146)	**0.56**	
ICL3	**bRh**	GQLVFTVKEAAAQQQESA	**18**, (224-241)		
	A2Ar	Insertion of T4 lysozyme			

β1AR	Insertion of T4 lysozyme
β2AR	Insertion of T4 lysozyme

3.2.2 A homology Model of β2AR from β1Ar

Given the success we had predicting a variety of GPCR loops in an imperfect environment, it seemed reasonable to approach a real homology modeling problem. Homology models present additional challenges of an imprecisely positioned TM bundle as well as several side chains potentially being in non-native conformations. Thus, for this study, we wanted the homology model to be close to the native structure in an effort to further validate that it is possible to accurately predict flexible protein domains in a slightly perturbed system. Table 3.2 presents the sequence identity of 36 GPCR pair combinations of bRh, β2Ar, β1Ar, A2Ar, and four newer structures that became available from the start of this project until the time we chose our first homology model attempt: CXCR4 (PDB ID code 3ODU) (73), D3 (PDB ID code 3PBL) (74), H1 (PDB ID code 3RZE) (75), and M2 (PDB ID code 3UON) (76). This table suggested that our first test case, based on sequence identity considerations, should be to build β2Ar from a β1Ar template, or to build β1Ar from a β2Ar template. Not only is their sequence identity percentage high, but we already knew that we could predict the loops of both proteins in less perturbed environments. The former was chosen. Between then and the time of this publication, there were four new structures published: S1P1R (PDB ID code 3V2Y) (77), KOR (PDB ID code 4DJH) (32), MOR (PDB ID code 4DKL) (78), and M3R (PDB ID code 4DAJ) (79). Thus, there are now three pairs within subclasses: β1 and β2 adrenergic receptors, M2 and M3 muscarinic acetylcholine receptors, and mu and kappa opioid receptors. All have high sequence identity percentages, and the latter two are sure to be useful for further validation in the future.

The TM bundle of the homology model of β2Ar based on the β1Ar template is very close

in structure to the native receptor. The RMSD between them is 0.88Å. The main deviations come from small kinks that distort the helices. These minor kinks can, however, have a significant influence on loop prediction. Even a very small change in a dihedral angle in the middle of a helix can, via a lever arm effect, lead to a significantly larger displacement of the terminal end of the helix. This effect is illustrated in Figure 3.3, which displays the native (purple) aligned structure of the C-terminus sides of TM helices 6 and 7 of β2Ar, and the corresponding residues of the homology model (green). The residues that the arrows are pointing toward have very well aligned backbones, but the angles between these and the next residue deviate slightly from the native structure, leading to the terminal residues of each homology model helix (upon which the loop prediction begins) being 1.09Å from the native for residue 299 and 0.72Å from the native for residue 304. Table 3.3 provides an analysis of the RMSDs of the flanking segments of the five loops of the homology model from the native structure, based on TM superposition. At these terminal residue displacement lengths, loop prediction remains successful. However, as these lengths increase, loop prediction gets more challenging and eventually impossible. The exact positioning of the loop stems is important for two reasons: first, it follows that the lowest energy (ie. crystal structure) loop must be slightly different from the native loop, and second, we have to recalibrate what we view as a successful prediction as viewed from the RMSD calculation. In earlier loop prediction work, an RMSD was considered excellent if it were less than 1.5Å for short loops (12 residues and less) and less than 2.5Å for long loops (13-20 residues). As our methods became increasingly better, we are now able to predict a loop with 20 or fewer residues with sub-1Å RMSD accuracy almost 100% of the time. A perfect backbone RMSD would be 0Å, when the loops align perfectly. The best possible prediction of a loop built on a homology model compared to the native loop cannot have an RMSD of 0Å, because the flanking residues

are not situated ideally (ie. as in the native structure) relative to one another. At best, one can recover the native loop shifted or stretched by the amount that the flanking residues differ, as compared to the native protein. In reality, the effect of shifted flanking residues, plus the additional perturbations throughout the entire protein, will lead to a predicted loop with at least minor variations in structure from start to end. The loops of a homology model thus have to be gauged by a softer standard than loops built on a native protein. While there is no perfect RMSD assessment, in general comparison of equivalent loops from a pair of different GPCR crystal structures of the same protein have RMSDs around 1Å-1.5Å with respect to each other. Such differences arise from various effects: alternative stabilization techniques that alter the positions of every atom, including the core region, different point group symmetries, and experimental error. Thus, for loop refinement of a homology model, a prediction that falls within this 1Å to 1.5Å range of the "native loop" captures the accuracy of another crystal structure if it contained similar perturbations throughout the entire protein. For very long loops this remains true as well. However, the great difficulty associated with predicting loops with such high conformational variability in a native GPCR is magnified in a homology model case. At this point it would be unrealistic to expect such great accuracy for these very long loops.

The results of loop prediction on the homology model of β2Ar built from the native β1Ar as the template are provided in Table 3.4. To better visualize these predicted loops, see Figure 3.1 panel c. The predictions of the short loops are within a tiny RMSD difference of the homology model as compared to the native structure. They also lie close to the expected lower-bound of accuracy for homology model loop prediction as discussed earlier.

Unsurprisingly, given the high sequence identity between β2Ar and β1Ar, the loops, with the exception of ICL2, are close in structure and thus the unrefined homology model already has

reasonably accurate loops. However, this did not mean that the predicted loop structures would maintain high fidelity to the native structure. We have significant evidence that even short loops that are being predicted in a system where the environment is perturbed can be reconstructed with large positional deviation from the native loop (data not shown). For some cases we have seen loops as short as 6 residues predicted with RMSDs over 5Å. This can occur even if the surrounding environment is reasonably close to the native. For the short loops, ECL1, ECL3, and ICL1, the predicted loops are restored with high accuracy and compare favorably to the loops obtained simply from the procedure used to build the homology model. Objectively, the predicted ECL2 is quite good by any measure, given that it is 26 residues long. It captures the correct overall position relative to the rest of the protein, as well as the correct folds. Nonetheless, it is significantly less in agreement with experiment than the homology model loop, which reflects the fact that ECL2 of β2Ar and β1Ar are extremely similar in structure and positional alignment in space. For every other target ECL2 built from β1Ar as a template (with the likely exception of β3Ar) an unrefined homology model would produce an ECL2 that bears little resemblance to the true structure. Thus, we consider this loop prediction successful: it provides good evidence that if we had built a model of β2Ar from a different template and the TM bundle were as accurate as the one obtained from the β1Ar template, we would still be able to arrive at a final predicted loop structure that has good fidelity to the native loop. Furthermore, for the portion of ECL2 that is most important for ligand binding—residues 191-196, C-terminal of the disulphide bridge—we obtain an RMSD of 1.50Å, a result that very likely falls within the required accuracy for ligand docking experiments (although this point needs to be established by doing such experiments explicitly). Figure 3.4 provides visualization of this region of the loop prediction compared with the native structure.

To predict ECL2 of β1Ar with a small helix in it, we used the same approach that we took in the past (including the perturbed native calculations). Again, we knew beforehand that ECL2 of β1Ar contains eight residues (PQALKCYQ) in the center of the loop that form a small helix, thus we guess that β2Ar might contain a helix in the same region. Therefore, we ran the ECL2 loop prediction calculation with and without a helical constraint on those residues. The loop predicted with the helical constraint was 28.93kcal lower in energy than the loop which did not specify a helix region. Four quadrant phase space partitioning was also used.

ICL2 presented a different challenge. 2RH1, the original and highest resolution structure to date of β2Ar, is in its inactive form and is stabilized by the insertion of T4-lysozyme in the location of ICL3. In 2RH1, this creates a structural change in ICL2. The same occurs in another crystal structure of β2Ar that is stabilized by T4-lysozyme (PDB ID code 3D4S). However, a newer structure of β2Ar in its active state, (PDB ID code 3P0G) does not have this problem, and the ICL2 of this structure contains a small helix. The various structures of β1Ar also contain ICL2s that have small helices, and the authors argue (41) that the conformation of ICL2 with the helix is representative of the physiologically relevant structure for all inactive βArs. In our original work, we predicted the structure of ICL2 using the 2RH1 crystal structure, including the T4-lysozyme. In this work, we remove the T4-lysozyme, but the helices remain in their crystallographic positions. This means that ICL2's contextual preference for the L-shaped strand like structure should remain, since the residues distorted by the T4-lysozyme remain in place, and we expect to be able to predict the loop close to the crystal structure. However, the homology model of β2Ar is based on the coordinates of β1Ar, thus we expect that the ICL2 that can be accommodated by this conformation of the local region would contain a small helical region. Other research (80) confirms the idea that the TM helices' conformation is the main

structural determinant of ICL2. Consequently, the best comparison for our loop prediction of ICL2 given the β1Ar template is with ICL2 of 3P0G, not 2RH1. In Table 3.4, the first line of RMSDs of ICL2 are of the homology model and our predicted loop versus ICL2 of 2RH1. The second line of RMSDs compare the homology model and predicted ICL2 with that of 3P0G.

The question that remains, of course, is just how close in position to the native does the homology model TM bundle have to be to refine loops of GPCRs as accurately as we are able to in this case. Evidence points to the idea that the local environment of a loop must have high fidelity to the native for loop refinement to be precise. We do not know, however, to what extent more distant regions can deviate from the native structure without incorrect long range energy calculations making accurate loop prediction impossible. Great care will have to be taken to answer these questions and develop new algorithms that capture the structure of the greater loop environment and eliminate clashes caused by imperfect backbone and side chain atoms throughout the entire protein.

Table 3.2 The Percentage Sequence Identity Between Pairs of GPCRs. The percentage sequence identity between pairs of GPCRs. The T4-lysozyme residues are not included in the sequence identity calculations.

% Seq ID	bRh	β2Ar	β1Ar	A2Ar	CXCR4	D3R	H1R	M2R
bRh	100	15	18	19	18	26	18	21
β2Ar	15	100	62	28	21	36	31	27
β1Ar	18	62	100	32	21	38	33	29
A2Ar	19	28	32	100	17	30	33	25
CXCR4	18	21	21	17	100	22	24	21
D3R	26	36	38	30	22	100	31	30
H1R	18	31	33	33	24	31	100	37
M2R	21	27	29	25	21	30	37	100

Table 3.3 The Average Cα Displacement Between Native and Homology Model Terminal Residues of β2Ar's TM Helices.

	ICL1 (TM 1,2)	ECL1 (TM 2,3)	ICL2 (TM 3,4)	ECL2 (TM 4,5)	ECL3 (TM 6,7)
RMSD (Å)	0.69	1.02	0.62	0.6	0.7

Figure 3.1 Visualization of loops. a. The extracellular loops of β2Ar. The red loop is the native ECL1 (residues Lys97 – Phe101), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Met171 – Asn196) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues Gln299 – Arg304), and the black loop is the superimposed predicted ECL3. b. The intracellular loops of β2Ar. The red loop is the native ICL1 (residues Phe61 – Thr66), and the blue loop is the superimposed predicted ICL1. The green loop is the native ICL2 (residues Ser137 - Tyr146), and the pink loop is the superimposed predicted ICL2. c. The extracellular loops of bRh. The red loop is the native ECL1 (residues Gly101 – Phe105), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Val173 – Asn199) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues His278 – Gly284), and the black loop is the superimposed predicted ECL3. d. The intracellular loops of bRh. The red loop is the native ICL1 (residues His65 - Thr70), and the blue loop is the superimposed predicted ICL1. The green loop is the native ICL2 (residues Cys140 - Gly149), and the pink loop is the superimposed predicted ICL2. e. The predicted extracellular loops of the β2Ar homology model superimposed on the native β2Ar. The orange helices represent the homology model, and the aquamarine helices represent the native β2Ar. The red loop is the native ECL1 (residues Lys97 – Phe101), and the green loop is the superimposed predicted ECL1. The blue loop is the native ECL2 (residues Met171 – Asn196) and the pink loop is the superimposed predicted ECL2. The yellow loop is the native ECL3 (residues Gln299 – Arg304), and the black loop is the superimposed predicted ECL3. f. The predicted extracellular loops of the β2Ar homology model superimposed on the native β2Ar. The orange helices represent the homology model, and the aquamarine helices represent the native β2Ar (PDBID 2RH1). The yellow helices represent native TM4 and

TM5 of β2Ar (PDBID 3P0G). The red loop is the native ICL1 (PDBID 2RH1) (residues Phe61 – Thr66), and the blue loop is the superimposed predicted ICL1 (on the homology model). The pink loop is the native ICL2 (PDBID 2RH1) (residues Ser137 - Tyr146), and the green loop is the superimposed predicted ICL2 (on the homology model). The yellow loops is the native ICL2 (PDBID 3P0G). The predicted ICL2 on the homology model aligns much better with ICL2 from 3P0G than from 2RH1.

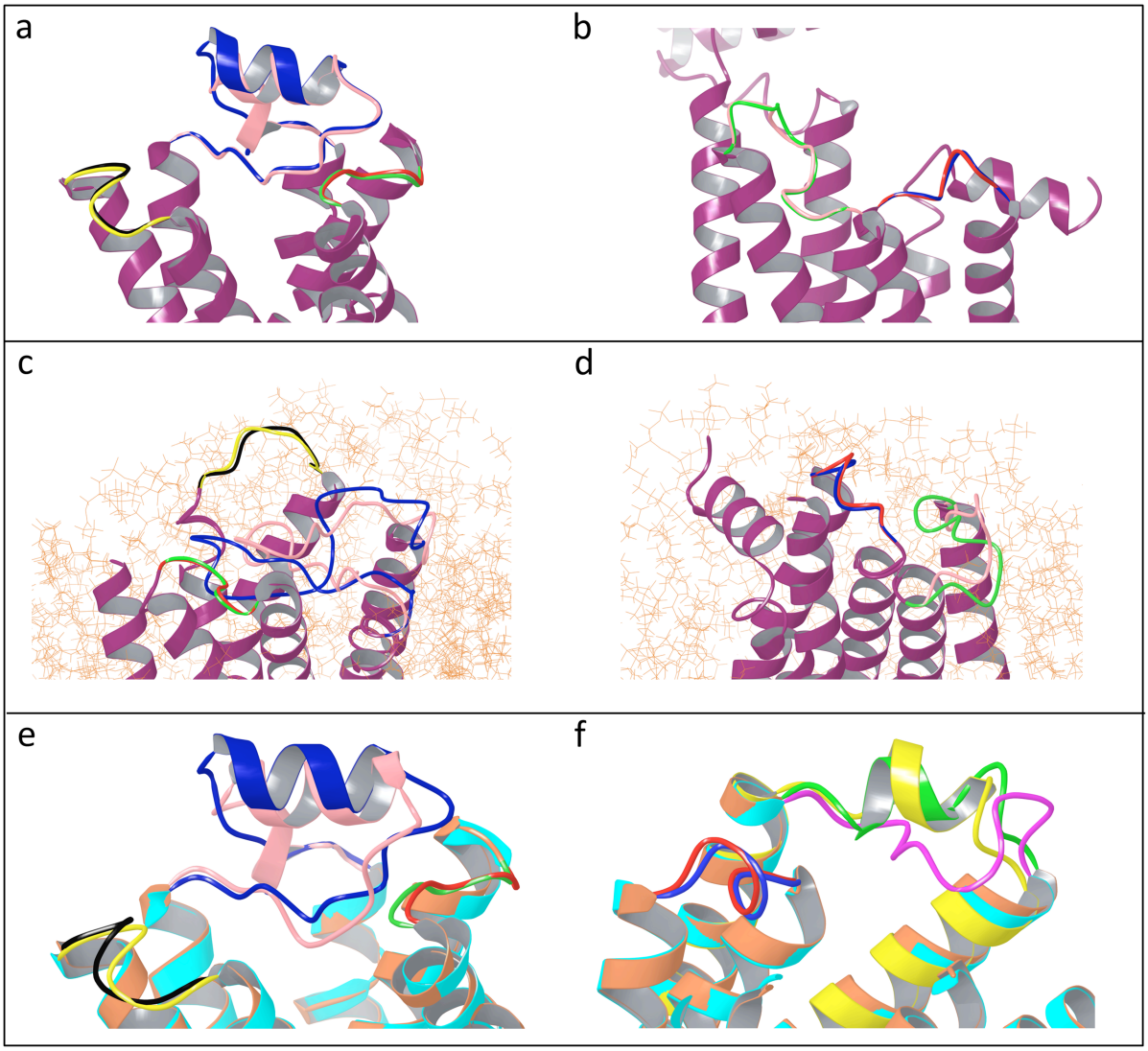


Figure 3.2 The native (purple) and predicted (green) ECL1 of A2Ar (residues Ser67-Ala73, between TM1 and TM2) surrounded by explicit membrane molecules (in red). The membrane molecules' are positioned such that there is unresolvable clash with the native loop, indicating a problem associated with equilibrating the bilayer without any knowledge of native loop position. Despite this problem, we are able to obtain a reasonable predicted loop structure for ECL1 when predicted with surrounding membrane molecules.

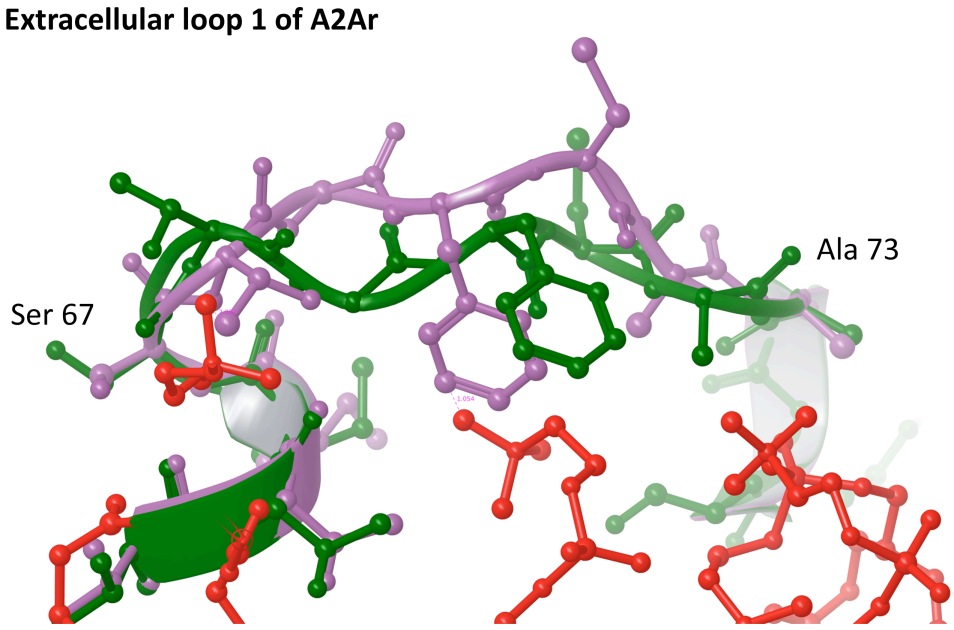


Figure 3.3 The C-terminus sides of TM helices 6 and 7 of the native (purple) and homology model (green) of β2Ar. Despite nearly perfect alignment where the arrows point, small kinks afterward lead to relatively large displacements of the helices' terminal residues, yet loop prediction remains successful. As terminal residue displacement between homology models and native proteins increases, accurate loop prediction becomes harder, and eventually potentially impossible.

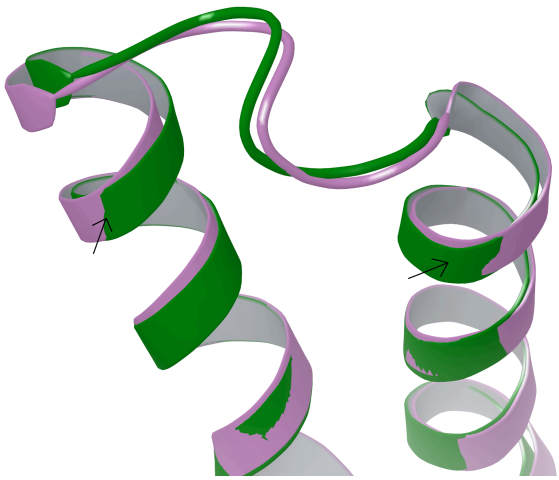
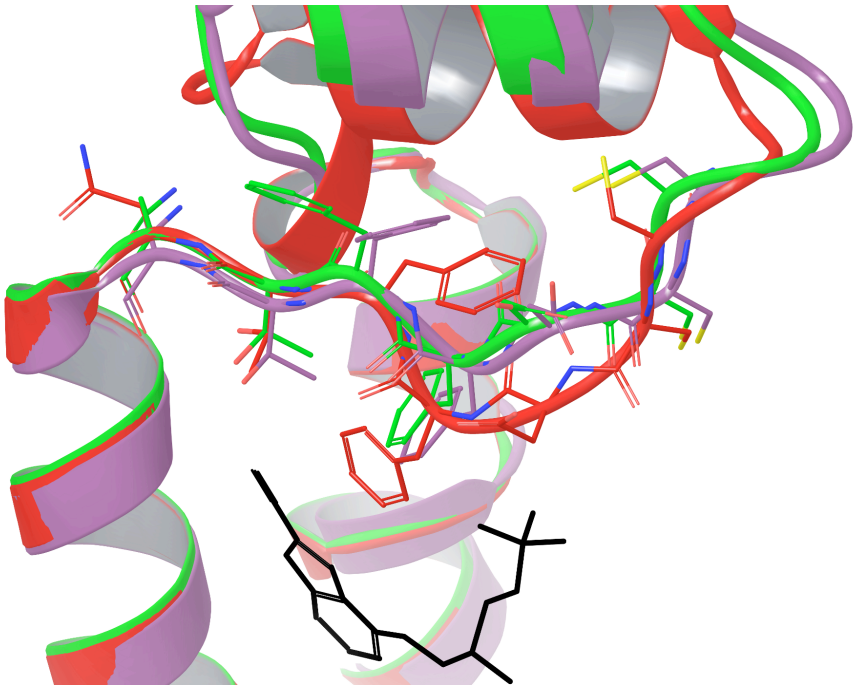


Table 3.4 The RMSDs of the loops on the β2Ar homology model. ICL2 contains two sets of RMSDs because the loop's structure is variable. a. The RMSD of the loops refined in the context of the homology model, as compared to the aligned native β2Ar structure. Note that for ICL2 the RMSD is calculated against two β2Ar structure: 2RH1 and 3P0G. b. The RMSD of the loops emerging directly from the homology model as compared to the aligned native β2Ar structure. Again, the RMSD of ICL2 is calculated against two β2Ar structure: 2RH1 and 3P0G.

	RMSD[a] (Å) of refined HM loops compared to aligned native β2Ar		**RMSD[b] (Å) of original HM loops compared to aligned native β2Ar**	
	PDBID 2RH1	**PDBID 3P0G**	**PDBID 2RH1**	**PDBID 3P0G**
ECL1	0.94		0.86	
ECL2	2.63 (1.50)		0.88	
ECL3	1.06		1.14	
ICL1	0.79		0.69	
ICL2	5.68	2.17	5.64	1.58

Figure 3.4 Residues 191-196 of ECL2 of β2Ar. The native protein is purple, the homology model, including the its original loop, is green, and the predicted loop is red. These residues are most important for ligand (carazolol is shown here in black) binding. The side chains are for the most part well aligned, although the predicted rotamer of residue Phe194 is closer to the native than in the homology model loop.



3.3 Conclusions

Loop prediction in imperfect environments is significantly more challenging than in the context of the native protein for two main reasons: deviations of atomic position from the native structure and new types of atomic clashes (arising from such deviations) that the energy function must pick out and penalize appropriately so that these incorrect structures are not propagated throughout the various stages of hierarchical methodology. Homology models represent an ultimate case of an imperfect environment, because every single atom is now in its non-native position. This causes a slew of issues to contend with such as the backbone of the protein core being kinked into an incorrect trajectory, side chains having inaccurate placement thereby blocking correct placement of other backbone and side chain atoms, and changes to the lowest energy structure of flexible loop regions itself. There are instances of loops (such as ICL2 of the βArs) that appear to be more influenced by the precise context of the TM structure than the sequences themselves. Even amongst homology models, however, there are varying degrees of difficulty. If the target and template align very well (which often corresponds to higher sequence identity), the resulting homology model will obviously be a much better starting point for loop prediction than if the target and template are highly divergent.

The results of this analysis of GPCR loop prediction with two different levels of imperfect environments—one where the TM residues are held fixed in the crystallographic positions and a real homology model where the TM domain is no longer exact, but still quite close—demonstrates that despite the complexities reiterated before, we are able to predict loops with high fidelity to their native counterparts. To the best of our knowledge, this is the first successful example of an RMSD validated, physics based loop prediction in the context of a GPCR homology model. To overcome difficulties that arise from non-native environments, we

used extra side chain sampling, explicit membrane calculations, and helical constraint methods. We also created a new phase space partitioning method that allows for increased, higher resolution sampling of a loop by limiting the position the central, closure residue.

Being able to predict loop structures one at a time, using atomic coordinates that are close to the crystal structure is necessary, but not sufficient, evidence to claim that we would be able to get results of similar quality for a homology model that contains a less accurate core region. Nonetheless, these results represent a very encouraging step forward in GPCR homology modeling loop refinement. One can imagine that for harder and more practical cases, increased sampling of the surrounding regions and further fine-tuned components of the energy function will ultimately allow us to build accurate GPCR homology models that would be of great importance to drug discovery initiatives as well as much basic science computational studies of understanding GPCR function. Lastly, our work, while currently tailored to GPCRs, is in no way limited to them and extends to all important protein families.

3.4 Methods

3.4.1 Overview of PLOP

One of PLOP's main functionalities is predicting loop structure from amino acid sequence, maintaining the rest of the protein fixed in its starting structure, whether it is the native crystal structure or a homology model. A single execution of PLOP generates thousands of loops that are clustered discriminated between by a physics-based energy function. The output is a couple of dozen distinct loop conformations that are ranked by energy. Loop prediction occurs in four stages: 1. buildup, 2. closure, 3. clustering, 4. scoring. We first briefly describe these stages, and then summarize the higher level hierarchical scheme that takes advantage of running multiple PLOP executions in parallel. Finally we discuss a new sampling method developed for

this project that divides sampling efforts into different regions of phase space. Figure 3.5 contains a flowchart outlining the major steps of single and full loop prediction, as well as the new phase space partitioning method to guide the reader through our loop prediction methodologies. For all calculations that include an explicit membrane or are done on a homology model, crystal symmetry information is not used. The other calculations do use symmetry information, meaning that crystal neighbors are included in the calculations. To ensure that they do not positively bias the result by blocking regions of space for loop buildup, thereby guiding the central asymmetric unit, all copies of the asymmetric unit are predicted simultaneously.

During the buildup stage, an initial set of right and left half-loop conformations are generated via a dihedral angle search through rotamer libraries. There are two sets of rotamer libraries: one containing (φ,ψ) angles representative of the Ramachandran plot for a single amino acid residue, and one containing (φ_1,ψ_1,ω, φ_2,ψ_2) dipeptide torsion angles, the latter arising from two sequential residues. In this study, the former is used for short (less than 10 residues) loops, while the latter is used for longer loops. Starting with the right and left loop stems, residues are added sequentially and terminate at the middle (closure) residue. Half-loop buildup starts at a very coarse resolution and decreases down to a lowest resolution of 5˚ (ie. measure of differences between rotamer states), until a pre-specified number of loop candidates are generated. The sheer number of half loops is reduced by means of a hard sphere steric clash check. It relies on a parameter called the overlap factor (*ofac*), which is the ratio of the distance between two atom centers and the sum of their atomic radii. As each residue is added the *ofac**atomic_distance (dist1) is compared to the user-specified *ofac*_cutoff*atomic_distance (dist2) for each atom pair between the residue and nearby atoms. As long as dist1<dist2, half-loop buildup continues. During the closure stage, pairs of half-loops are screened so that they have closure C_α atoms

within 0.5Å of each other, a closure N-C_α-C angle near the ideal value of 111.1˚, and no major clashes. These are the original set of loop candidates. To reduce structural overlap between loop candidates and reduce the required computer time for the algorithm, the loops undergo a modified K-means clustering algorithm that clusters loops by RMSD. The loop closest to the center of each cluster is then sent forward to the last stage of loop prediction, optimization and scoring. In this final phase of loop prediction, side chains are optimized using side chain rotamer libraries that have 10˚ resolution (described in more detail later), and the entire structure is minimized. The energy of each of the final set of loop candidates is calculated with our newest energy function using an implicit solvent model (28), and the lowest energy loop is the final prediction of this single PLOP execution. The energy function is based on the OPLS all atom force field for bonded and nonbonded terms, coupled to a generalized Born based continuum solvation description. Inclusion of a hydrophobic term(81) optimized to reproduce protein-ligand binding affinities (as opposed to the usual fitting to solvation free energies of small linear hydrocarbons), use of a variable internal dielectric to approximately represent enhanced polarization effects arising from charged side chains, and empirically optimized (but physics-based in motivation) hydrogen bonding, π-π interactions, and self-contact interactions corrections make the energy model distinct from others. The excellent performance of the model for both single side chain prediction and loop prediction, which represents a substantial advance as compared to previous efforts, is described in detail in ref. (28).

As a single execution of PLOP creates tens of thousands of loops, to better sample phase space, we use a hierarchical scheme that contains several stages. Each stage involves multiple loop predictions run in parallel, whose starting points come from the lowest energy predictions from all previous stages combined. At each stage, varying input parameters are used, leading to

further conformational sampling with a slightly different focus. The first, or *Init*, stage, includes five PLOP jobs, each with a different *ofac* cutoff: lower values correspond to higher tolerance for atomic clashes. For this study, we used low *ofac* cutoffs: 0.30, 0.35, 0.40, 0.45, and 0.50. The 25 lowest energy structures from all of the *Init* stage jobs are then sent onto the first refinement stage, where loop buildup occurs as described before, but a 6Å Cartesian constraint is placed on the C_α atoms with respect to the structure imported from the first stage, thereby refining the structures already in a low energy well. Short loops then undergo a second refinement stage with a 4Å C_α constraint; longer loops first go through a series of fixed stages before the fine-tuned refinement. Fixed stages hold still the positions of the terminal residues of the loop of interest and predict only the shorter, interior loop fragment. For example, during the first fixed stage, one residue is held fixed, either the left terminal residue (and the next through last terminal residue are predicted), or the right terminal residue (and the first terminal through second to last residue are predicted). The 20 lowest energy structures from this stage are passed onto the second fixed stage, where the first two, last two, or first and last residues are held fixed in their previous positions. The lowest energy structures from the second fixed stage are passed onto the third fixed stage, and so on. Thus, all different combinations of the total number of terminal residues that can be held fixed are tried to increase sampling, while focusing loop sampling on smaller and smaller loops regions. For the long second extracellular loop of GPCRs, we used 10 fixed stages, as this was demonstrated to be an effective number of such calculations in ref. (28).

For loops that contain helical fragments we have developed a method documented in detail in ref. (29). This new method incorporates a different dipeptide library that contains coupled dihedral angles often found in helices. The helical portion of the loop is built up with this special library, thus ensuring that a helix forms in the loop. However, during the

minimization procedure, the helical region can unravel to give a lower energy loop structure. It is not the case that the structure containing a helix will always have a lower total free energy in the continuum solvent model than a structure without a helix. The helical structure can have unfavorable side chain interactions or poor solvation of polar or charged side chains, and the energy model, based on data accumulated to date, performs remarkably well in its ability to make robust structural predictions.

The overall algorithm is also slightly tweaked such that the closure residue is not contained by the helix. To use this algorithm, the user has to have an initial guess to specify the helical bounds, either by secondary structure prediction or by a homology modeling approach. In this study, we use the latter approach for ECL2 of the β adrenergic receptors: because β1Ar has a relatively high sequence identity to β2Ar, it is only logical to try building ECL2 with and without a helix. In such an approach, we also carry out a separate calculation in which no helical bounds are mandated. The final predictions of the two calculations are compared, and one with the lowest energy is selected as the final answer. This approach permits unbiased comparison of alternative structures, and to date has proven highly successful in treating helix containing loop regions for a wide range of test cases, while still recovering normal loop prediction when that is the correct result.

3.4.2 Phase space partitioning

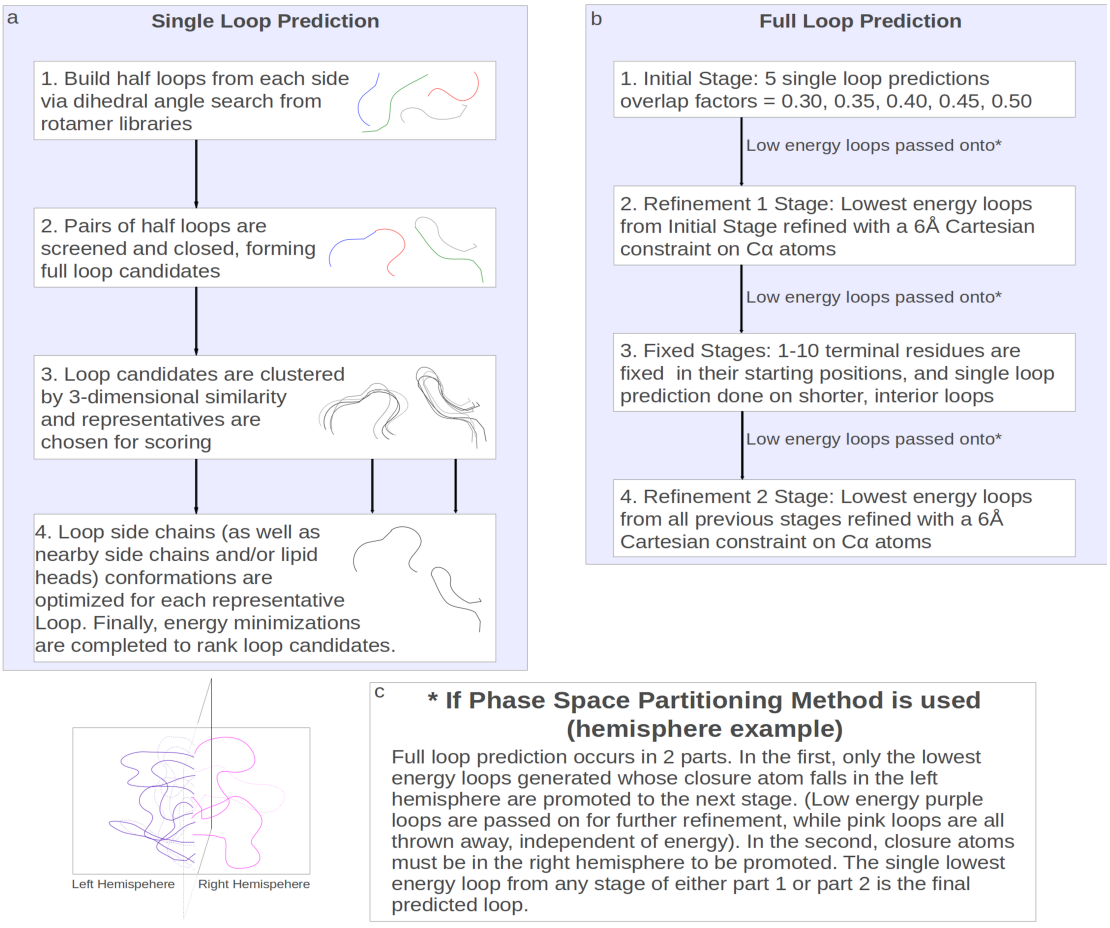
Because imperfect environments present extra sampling challenges, and the default sampling methodology described in ref. 6 was not able to identify the correct solution in the long loop cases tested here (data not shown) we developed a new method for extended sampling that we call the phase space partitioning method (PSPM). The basic idea is to partition the phase space into multiple regions and constrain the loop sampling in each PLOP run into one of these

regions. The combination of multiple sub-region sampling completes the sampling of the whole phase space The implementation relies on a new screening term, based on the closure atom, that checks where this atom is in the targeted sub-region of phase space. If the closure atom is located in the targeted region of phase space, then the built up half-loop is accepted, if it is in a different (non-targeted) region of phase space, the half-loop is rejected. To ensure unbiased sampling, phase space is divided into, for example, hemispheres or quadrants. For the hemisphere case, complete loop buildup occurs across two PLOP executions. To generate the phase space partitions, first a vector is defined between the Cα atoms of the starting and ending residue of the loop. The cross product between this vector, *vec1*, and the Cartesian basis vector (0,0,1), *vec2*, defines a normal vector, *vec3*, perpendicular to the plane defined by *vec1* and *vec2*. Similarly, the cross product between *vec1* and *vec3* defines a normal vector perpendicular to the plane that *vect1* and *vec3* lie in. As each rotamer of the closure residue for each left and right half-loop is tried, its atoms are screened to be in an appropriate half or quadrant of phase space, as divided by these vectors, and only structures that fall into the allowed region are kept as half-loop candidates for forming full loops. This increases the number of loops generated whose closure residue falls in the allowed region for a given PLOP execution. Because more half-loops are rejected during the buildup, each PLOP job can go to higher sampling resolution and thus achieve more complete sampling in the targeted region of phase space. Additionally, by ensuring that each targeted sub-region of phase space is sampled individually, the entire region of phase space is sampled more evenly. The principal benefit of the approach is manifested when initial loop scoring funnels candidates into a small phase space area, even if the correct loop structure occupies a different area, because the correct loop conformation requires more extensive sampling before any competitively low energy structures are generated. This is most likely to be

a problem for super long loops (such as the ECL2 loop in GPCR structures), and this is where the PSPM algorithm is deployed in the current work. The other GPCR loops are sufficiently short that they do not require this more computationally intensive technique.

The new screening element first occurs in the *Init* stage. For the hemisphere case, the total number of PLOP executions doubles, meaning that for each of the five *ofac* cutoffs tested, there are two associated PLOP jobs. For each *ofac* cutoff, the first PLOP job generates loops which contain closure atoms only in one hemisphere, while the closure atoms of the loops produced by the second PLOP job lie in the other hemisphere. For the quadrant case, there are 20 total PLOP jobs run (four for each *ofac* cutoff). For this study, the 25 lowest energy loops are passed onto the first refinement stage, where they are subject to a 6Å constraint on the Cα atoms. The 20 lowest energy loops are then passed onto the first fixed stage. Because the shorter fragments of the loop that are predicted during the fixed stages are free to move around space without any Cartesian constraints, we again divide phase space into equal sized volumes and sample loops within it. The final refinement stage places a 4Å constraint on the C_α atoms of the 10 lowest energy loops from all previous stages combined. The lowest energy loop after this last stage is the final predicted loop.

Figure 3.5 Flow charts and illustrations of loop prediction methodologies. a. A flow chart describing the 4 main steps of single loop prediction: buildup, closure, clustering, scoring. In step 1, half-loops are built, in step 2, half-loops that can meet in the middle are closed, in step 3, similar loops are clustered, and in step 4, representative loops are scored. b. A flow chart describing the various stages of full loop prediction. c. Visualization of the phase space partitioning method, using hemispheres as the example. Two full loop predictions are run. In each one, loops are promoted only if their closure atom falls in the prespecified hemisphere. The lowest energy loop coming from both full loop predictions is the final predicted loop.



3.4.3 RMSD calculations

We use root mean squared deviation, or RMSD, to gauge the accuracy of loop prediction. We calculate RMSD by superimposing the protein backbone, except for the loop of interest, of the native structure with that of the model protein onto which the loop is being built. The coordinates of the N, C_α and C_β atoms of the predicted and native loop are then used to calculate RMSD. Every RMSD cited in this paper is calculated in this way.

3.4.4 Loop prediction with surrounding side chain optimization

Because all of the structures used in this study have imprecise regions in addition to the target loop to be predicted (either other loops or, in the case of the homology model, in principle the entire protein), we employ another form of extended sampling in which additional side chains on the protein body within 7.5Å of the loop are sampled and optimized simultaneously (62). This is accomplished via an iterative optimization of side chain conformations that includes this expanded list of side chains during the scoring of the loop candidates. The loop side chains are optimized first, followed by the surrounding side chains. Each stage throughout hierarchical loop prediction starts with the optimized side chain formations garnered from the previous stage. The side chains are energy-minimized simultaneously to remove steric clashes. Optimization occurs in an iterative fashion in which each side chain is sampled, and the lowest energy rotamer state (in the context of the rest of the side chains' state) is picked. Convergence is achieved when for less than 5% of the side chains, a lower possible energy rotamer is found.

3.4.5 Explicit membrane calculations

As described in ref. (30), to include an explicit membrane into a GPCR loop prediction, we first equilibrate the membrane using molecular dynamics (described below). The explicit lipid molecules serve to prevent 1) loop prediction from proceeding in physically impossible

locations, and 2) electrostatically highly unfavorable events from occurring, such as burial of a charged residue in the hydrocarbon region of the membrane. They also provide generally better energy assessments, as loops that interact with the membrane are coupled to non-solvent atoms, and the dielectric must thus be different. Once the membrane is equilibrated, a loop prediction occurs as described before, except that now the positions of both the loop side chains and the surrounding lipid heads within 7.5Å of loop atoms are optimized. This is accomplished in a similar way that nearby side chains included in the loop prediction are optimized: each lipid head is opimized one at a time by sampling 3 key torsional angles at 10˚ resolution, and the lowest energy conformation in context of the rest of the lipid heads (updated for each new lipid being sampled) is picked until convergence is reached. In this way, the lipid heads are energy-minimized simultaneously to prevent clash. The new optimized side chain and lipid head orientations are used as the new starting positions for each stage of loop prediction. This procedure prevents the specific orientation of the lipid molecules from incorrectly biasing loop prediction by giving the lipid heads flexibility. The various loop plus lipid positions are scored using the all-atom energy function within PLOP.

3.4.6 Molecular dynamics simulations

The explicit solvent MD simulations were run with Desmond v3.0, available from Schrodinger modeling suite Maestro Version 9.3. The system was prepared with Desmond System Builder. First, the membrane position was obtained from the OPM (82) database. For 1U19 (bovine rhodopsin), the OPM database did not have the corresponding PDB ID, so the membrane position was taken from 3CAP (39) (squid rhodopsin) instead. Then, explicit membrane molecules were used to fill the space between the upper and lower bounds of the GPCRs, as defined by the two membrane planes. The lipid membrane molecules were positioned

according to this OPM membrane thickness, and the polar heads are outside the membrane planes, while the aliphatic tails are inside the planes. The orthorhombic boundary condition is used. Any lipid membrane molecules that overlapped with the protein were removed. 10Å of SPC water molecules were also included above and below the protein-membrane system. Any waters that ended up inside the lipid bilayer were removed. We used POPC for the bRh and A2Ar systems, which simulates the membrane properties well (83). The lipids and proteins were parametrized with the OPLS 2005 force field (33, 34), and the water model is SPC. The system net charges were neutralized by adding Cl^- and Na^+ ions. The solvent and membrane relaxation and equilibration was done using the Desmond Utility "multisim" workflow. More specifically, the system was first minimized to the gradient of 10 kcal/mol/Å with maximum 2000 steepest descent steps. A harmonic restraint with the force constant of 50.0 kcal/mol/$Å^2$ was applied to all protein heavy atoms. Then the system was gradually heated from 0 to 323K in a span of 0.06ns, followed by 0.3ns NPT simulation to allow the equilibration of the solvent and lipids. The harmonic restraint described above was applied to protein heavy atoms during this stage. Finally, a 0.6ns NPT simulation at 323K was run while the protein restraint was reduced from 50.0 to 10.0 gradually. The final structure was collected from the end of the simulation. For the native bRh and A2Ar structures, the membrane equilibration process only included the positions of the TM bundle residues. For the homology model, the entire model was used.

3.4.7 Homology modeling

All GPCR structures cited throughout this paper were retrieved from the Protein Data Bank. The sequences were extracted using the Multiple Sequence Viewer in Maestro 9.3 and were aligned using ClustalW (84). Manual refinement was done to correct for the alignment of loops as well as to mitigate unphysical insertions and deletions. The homology model was

generated using Prime (85) with its default settings, based on the pairwise sequence alignment in that program. For the β2Ar/ β1Ar target/template pair, the ligand that binds to the 2RH1 template (carazolol) was shape aligned using the flexible alignment tool in Maestro 9.3 with the native 2VT4 ligand (cyanopindolol), and this positioning was used for all extracellular loop predictions performed on the homology model. Loops with close proximity to the cocrystallized ligand have their structure affected by it, due to both steric and electrostatic effects. Therefore, the only fair way to compare a predicted loop structure with the native is to include the ligand. While we could have used carazolol for the loop predictions on the β2Ar homology model, using a β2Ar ligand made more sense. Since the two receptors are very similar, we chose to align the two ligands in the same part of the binding pocket. In a case where one is less sure where the ligand is most likely to bind, one could either choose to leave the ligand out, or dock the ligand to find its lowest energy position and conformation.

Chapter 4

Prediction of long loops with embedded secondary structure using the protein local optimization program

4.1 Note

The majority of this work was done by my wonderful colleague, Edward Miller. When Ed came to the lab, I had already started on fleshing out the loop-helix-loop loop prediction technology that was started by Suwen Zhao and Kai Zhu. I suggested that he work on the project, and helped him come up with the overall plan. What emerged was a very detailed and thorough paper that highlights the best of PLOP's predictive abilities to date.

4.2 Introduction

Continual advances in loop prediction have yielded accurate modeling from twelve-residue loops (86) up to loops as long as twenty residues (28). These methods have managed to achieve near-atomic accuracy performing loop prediction in the presence of the crystal structure environment – a necessary, but not sufficient condition for realistic homology modeling.

Historically, loop prediction was first approached analytically by Go and Scheraga (87) in 1970. Demonstrated was the ability to predict, by solving a set of equations, the conformation of peptide fragments containing up to six rotatable torsions. This analytical method was updated 21 years later by Palmer and Scheraga (88). Here, the authors relax constraints on the original formulation by permitting each residue in the loop to adopt independent bond lengths or bond angles. However, the analytical method still remained limited to six torsion angles - three residues assuming the backbone ω torsion remained fixed. To accommodate larger loops, Palmer and Scheraga extend the method by permitting additional torsions, beyond the six that can be analytically determined, so long as they are independently set prior to the calculations. Thus,

their method requires that the algorithm be repeated numerous times over a conformational search of these additional independent torsions. Hence, for larger loops combinatorics must be considered.

Moult and James in 1986 proposed one of the first combinatorial searches through a discrete set of torsions (89). Here, the authors described the use of a systematic search through torsion angles obtained from a Ramachandran plot. For loops as small as five residues, their method yields about 10^{10} conformations, already an intractable number. To cope with the combinatorial explosion the authors, employ the use of rules and filters to restrict and prune the number of conformations to a manageable subset before performing more expensive scoring. Loops are scored with using a simple pairwise electrostatic energy function and a surface area based hydrophobic term.

Later methods vary in both the sampling rules and scoring function. Bruccoleri and Karplus in 1987 released CONGEN, from which our algorithm draws some similarity. There the authors use the CHARMM energy function (60) to score loops. In 1992, Bassolino-Kilmas and Bruccoleri advance CONGEN to permit directed loop buildup which takes into account information from partially built structures (90). In 2003, DePristo et al. (91) and de Bakker et al. (92) use the AMBER forcefield (93) and Generalized Born solvation model (94) for scoring loops. Loop buildup is performed using, among other modifications, a fine-grained torsion library that is residue-specific. Like CONGEN, our work draws similarities to this last method (86). We note that this historical review is not exhaustive but is intended to highlight the origins of loop prediction as it relates to this work.

In general, the use of combinational exploration of torsion space for loop buildup has within it two sub-problems, sampling problems where coping with the combinatorics of loop

buildup requires the development of clever pruning strategies, and energy problems where the minimization, scoring and ranking of the resultant loops must be computationally affordable yet accurate enough to identify the best conformation among those produced.

Throughout the literature, the functional definition of a loop has been a local segment of the protein that is free of secondary structure other than, perhaps, three-residue 3^{10} helices, but lies between large, likely well-conserved, secondary structure elements. Indeed, initial homology models are often constructed on the assumption that secondary structure elements are conserved between the template and the target (95). However, this loop definition has not always been strictly followed. Notable cases of loops containing secondary structure are the ECL2 loops of human β2-adrenergic receptor (42) and turkey β1-adrenergic receptor (41), both G-protein coupled receptors (GPCRs). These loops are actually loop-helix-loops (LHLs) containing an eight-residue α-helix. Spinach Rubisco is another example. The active site is composed of a highly conserved α/β barrel. Lying between each α/β pair are loops, of which loop 5 contains a five-residue α-helix and two residues that form part of βF, a β-strand external to the active site, and loop 8 which contains a four-residue α-helix (96).

Recent attempts have been made to model the GPCR LHLs and have been met with significant success reaching an accuracy as high as a 1.59 Å RMSD (30). As the method we provide here exists along a continuum of protein structure prediction methods, one that shares significant applicability to secondary structure-free loops, we retain the loose definition of the word ‘loops’, and here refer to loops as a region of the protein that may contain secondary structure but is flanked by even larger secondary structure elements. Presented in greater detail below is a precise definition, which was strictly enforced, to select a set of test of cases.

Throughout the literature, predictions performed on loops containing secondary structure

are scant. Zhu, Xie and Honig presented a refinement protocol that addresses loop-helix-loops and loop-hairpin-loops, referred to more generally as protein segments in the paper, using a knowledge-based potential(97). What is explored is the refinement of these segments, rather than the prediction of the segments *de novo*. Consequently, the success of their refinement is dependent on the difficulty of the initial structure. For hairpins and loop-helix-loops, close to 70% of their refinements yield predictions with an RMSD of 2.0 Å or better. In these cases, the secondary structure elements are kept fixed with their native torsions and moved as a rigid body. However, as our method discussed in this paper is independent of the conformation of the input loop (although it is dependent on the conformation of the surrounding environment) results cannot be directly compared.

Alternatively, Rohl *et al.*, described *de novo* loop construction using the Rosetta algorithm (98). Included in their test set are predictions of ten loops, referred to as structurally variable regions, of 13 to 34 residues in length. These predictions were done in the crystal structure environment and do include loops containing secondary structure. Although some of the members of their test set include, for example, loop-helix-loops, only ten cases were done in the context of the native protein – too few to permit comparisons between our method without relying on anecdotal information. Instead, the authors concentrate on the more ambitious task of loop prediction in an unrefined homology model. Finally, we note in a previous study, our attempt to address the challenges of helix packing (99). In Li *et al.,* we explored placement of a helix in a loop-helix-loop but treated the helix as a rigid body. Although the method relies on prior knowledge of the presence of a helix, for large helices, this is not unreasonable, as is stated above, because significant segments of secondary structure tend to be conserved across homologous structures. Indeed, the smallest helix considered in this study was eight-residues.

To the best of our knowledge, no studies have been performed that systematically address the challenges of *de novo* prediction of loops containing secondary structure, particularly for cases when *a priori* knowledge about the presence of small secondary structure is noisy at best. As loop prediction matures to accurate prediction of larger and larger loops, it becomes awkward to exclude cases of secondary structure-embedded loops. In this work, we propose a method to predict long loops containing possibly multiple helices or a hairpin. Our initial test set is composed of loops containing between 8 and 17 residues. The secondary structure length explored ranges from 3 to 13 residues, although in principle, prediction of loops containing larger secondary structure segments remains tractable.

For loop-helix-loops, we constructed a separate dihedral library taken from a non-redundant set of high-resolution Protein Data Bank (25) structures containing α-helices. The user is required to specify which residues this helical dihedral library is to be applied to, termed the helical bounds. Results with exact helical bounds taken from the crystal structure were used as an initial validation. More relevant to actual structure prediction and refinement, we then concentrated on accurate loop prediction using helical bounds supplied by either sequence-based secondary structure prediction algorithms or previous loop predictions performed without the use of our helical dihedral library. That is, in many cases, nascent helices were predicted without supplying any expectation of a helix. This suggested a propensity for this loop to include a helix and allow us to repredict the loop using our helical dihedral library. Throughout all sampling methods explored, what remains crucial is that purely from our energy model, we are able to pick out the loop with the lowest, or near lowest RMSD relative to the native structure. Finally, for loops containing either helices or hairpins, we explored loop reprediction in a perturbed local environment, similar to an environment encountered in full homology models, although without

deviations of the backbone from the native structure, and established success in restoring the native loop conformation. The results are generally satisfactory with loop-helix-loop predictions from imprecise helical bounds routinely reaching sub-Ångström RMSD and hairpin predictions reaching similar atomic accuracy.

4.3 Materials and methods

4.3.1 Selection of test cases

All PDB structures that were available as of August 30, 2010 were searched. A global criteria was used to select structures that satisfy the following properties:

1. A sequence identity between any two proteins must be ≤ 50%
2. Only crystal structures were selected
3. The resolution of the crystal structure must be < 2.0Å
4. Structures reporting only Cα coordinates were excluded
5. A minimum R_{work} of 0.25 was enforced.
6. The pH of the crystal structure was restricted to lie between 6.0 and 8.0.

The exclusion of proteins due to sequence identity was performed using the PISCES web server (100) (http://dunbrack.fccc.edu/PISCES.php). Loops were selected using a local criterion that satisfies the following:

1. The average temperature factor of atoms within the loop must be ≤ 35.
2. The real-space R-factor (101) of any residues in a selected target loop must not be greater than 0.200.
3. All residues within the loop or interacting with any residues within the loop must be free of alternate conformations.
4. To reduce effects due to loop-ligand interactions, the minimum distance between any

loop atom and any atom as part of a neutral ligand must be > 4 Å. For charged ligands, this cutoff is increased to 6.5 Å.

The real-space R-factor was found by reference to the Uppsala Electron Density Server (http://eds.bmc.uu.se/eds/) (102). The above criteria are similar to what was used to create test sets in our past publications.

4.3.2 Identification of secondary structure-containing loops

In our most recent publications, loops were defined as being a segment of the protein absent of secondary structure (58). To identify loops containing secondary structure, an alternative definition was proposed. For loops containing secondary structure, the loop must be bounded by a span of secondary-structure larger than the greatest contiguous span of secondary structure within the loop. For example, if a loop contained, at most, a six-residue α-helix, then flanking the loop must be residues that are a part of a secondary structure element of at least seven residues in length. Furthermore, the first and last residue of a loop must also not display secondary structure. Assignment of secondary structure on a per residue basis was done using the DSSP program (103).

A loop was defined as a loop-helix-loop only if there were no other types of secondary structure present other than turns and helices (including 3^{10} and α-helices), i.e. any loop containing both β-bridges and helical residues was discarded from this study. A total of 35 loop-helix-loop regions were identified which were either 16 or 17 residues in length in all. This loop length was chosen to select cases that were considered sufficiently difficult to demonstrate the efficacy of our approach. In our previous publication, loops free of secondary-structure were successfully predicted up 17 residues in length(104).

For loops containing β-hairpins, it became necessary to distinguish between a β-hairpin

and a segment that is part of a larger β-sheet. To make such a distinction, the following criteria were used:

1. The loop must contain the secondary structure pattern strand-turn-strand.
2. However, the turn residues need not be immediately adjacent to a strand residue.
3. The loop must be free of helices.
4. The strand residues comprising part of the pattern in criterion 1 must be forming backbone hydrogen bonds only to other residues within the loop.
5. The hydrogen-bonding pattern must be anti-parallel.

For hairpins, requiring loops be either 16 or 17 residues in length yielded too few test cases. Thus, a loop was accepted so long as it was not greater than 17 residues. A total of 41 cases satisfying the above hairpin criteria were identified.

4.3.3 Single loop prediction

Single loop prediction is performed through individual runs of the Protein Local Optimization Program (PLOP). Briefly, PLOP operates through four stages: buildup, closure, clustering, and scoring. Full details can be found in Jacobson *et al*. (27), however, the salient features will be presented here and the modifications of the PLOP protocol utilized in this work will be described.

Loop buildup is begun with a backbone dihedral angle library constructed from rotamers frequently observed in crystal structures. Initially, the library contained a set of dihedrals on a single amino acid basis(86). As larger loops were explored, efficient exploration of conformational space dictated the use of a dipeptide dihedral library (71). In this approach, a library is constructed from each of the 400 (20 x 20) possible dipeptide pairs and used in a sequence specific manner during buildup. For example, a loop containing an arginine–alanine

dipeptide would explore sampling from a different rotamer library than an arginine–valine dipeptide. This implicitly treats the individual amino acid torsions as coupled.

In helices, the backbone torsions are highly coupled to form the necessary hydrogen-bonding network. It was therefore natural to extend the use of a dipeptide dihedral library to exploit coupled backbone torsions across the four residues, or greater, of an α-helix. As such, for residues considered to be helical, a separate n-residue α-helical library was used for loop buildup, where n is four or larger. The aspects of this α-helical library are discussed in greater detail below. In β-hairpins, non-local torsional coupling is present and so to enforce torsional coupling during loop buildup would heavily constrain both the coupled, hydrogen-bonding residues, as well as the intervening turn residues. Although such an approach may still be fruitful, we found that for β-hairpins, our previous dipeptide torsional library was effective and so we did not explore further the use of an alternative β-hairpin library.

Loop buildup is performed simultaneously from both ends of the loop up to the C_α atom on the closure residue. In our prior publications, the closure residue was simply picked as the midpoint of the loop. For the loop-helix-loops described in this work, the closure residue, shared by both halves of a loop, cannot be permitted to bisect a helix. As is described further below, the helical library is based on the construction of entire helices, and not helical fragments. If the closure residue of the loop were a part of a helix, the helix would be split between both halves of the loop. Thus for this work, we were forced to alter the designation of the closure residue. The closure residue is initially set with the equation

$$C_{a,\,closure} = N_{term,\,LHL} + \left(Length_{LHL} - 1 \pm Length_{Helix}\right)/2$$

where $+$ is used when the C-terminus loop is the longer loop and $-$ for when the N-terminus loop is longer or if both flanking loops are of equal length. $N_{term,LHL}$ refers to the residue number of

the N-terminus of the loop-helix-loop. Should the closure lie adjacent to the helix, the closure residue is shifted one residue further away from the helix. This is to afford extra flexibility to the residues that precede loop closure.

Clarifying by example, consider the LHL predicted in PDB 1BKR (Figure 4.1). Predicted was the 17-residue loop-helix-loop from G75 – D91 containing a 4-residue alpha helix from P82 to I85. When predicting this loop without the helical library the closure residue is at the midpoint of the LHL, residue 83, highlighted in white in Figure 4.1. This residue intersects the helix and so cannot serve as the closure residue when employing the helical dihedral library from segments 82-85. Application of the above equation places the closure residue adjacent to the helix at residue D81, but for further flexibility, the closure residue is assigned to be residue L80 on the N-terminus loop, two residues away from the start of the helix. As in our previous work, the Cartesian positions of the two closure C_α atoms are averaged and the remaining atoms of the loop backbone are generated using standard geometry algorithms to close the loop.

During loop buildup, nascent loops undergo preliminary screening through the use of a parameter termed the overlap factor (*ofac*). The *ofac* is defined as the ratio of the distance between two atom centers to the sum of their atomic radii. A lower *ofac* cutoff allows for a higher overlap between the van der Waals radii. If during loop buildup, a backbone atom is placed with a smaller *ofac* than permitted by the threshold, then that candidate loop is discarded.

Three additional screens are used to reject unreasonable loops early in their construction: For the current residue(s) being predicted, there must exist at least one acceptable side-chain conformation, based on sampling a 30º side-chain rotamer library.

The loop must not travel further than 6.32 Å away from every C_α atom in the protein. This is an empirically determined value and is meant to reject loops that fail to form contacts with the rest

of the protein.

The distance between the latest residue predicted and the closure residue must be less than a threshold beyond which closure is not considered possible. For example, a statistical analysis of a set of >500 proteins found that the maximum C_α - C_α distance that can be spanned by four residues is 13.97 Å.

Full details of these screening methods are given in Jacobson *et al* (86).

An additional screening method is also employed to enforce broad sampling of conformational space. During loop buildup via single dihedrals, all pairs of states must obey the relationship $\Delta\phi^2+\Delta\psi^2>R_{eff}^2$, where R_{eff} is the "effective resolution" of (ϕ,ψ) space. The effective resolution is adaptively set during loop buildup. The total number of loop candidates is constrained to lie between a minimum of 512 loops up to a maximum of 10^6 loops. This constrains the number of loop candidates to a tractable size. We achieve this by initially setting the effective resolution to a coarse value of 300º and then gradually improve the resolution to finer values down to a minimum of 5º (the resolution limit of the dihedral library). For loop buildup using the dipeptide dihedral library, the effective resolution relationship becomes:

$$\Delta\phi_1^2+\Delta\psi_1^2+\Delta\omega^2+\Delta\phi_2^2+\Delta\psi_2^2>R_{eff}^2 .$$

Loop buildup using the helical dihedral library did not utilize any effective resolution relationship. Principally, this was because the size of the helical dihedral library is significantly smaller than the single peptide or dipeptide dihedral library. Due to a "lever effect", a small change in the dihedrals at one end of a helix can significantly alter the coordinates of the opposite end of the helix. This effect becomes more dramatic for larger helices. To exclude what few candidate loops are produced during buildup because of a resolution cutoff would be to ignore this lever effect. Greater detail about the construction and composition of the helix

dihedral library is presented below.

To prevent expensive optimization of similar loop candidates, the k-means clustering algorithm(105, 106) is employed and only one representative loop per cluster is passed onto side chain sampling and optimization. The number of clusters is set to be four times the number of residues in a loop, excluding residues initially flagged as helical during input to loop prediction, up to a preset maximum of 50 clusters. The number of clusters determines the number of representative loops passed onto side chain sampling/loop optimization and is empirically set to balance the conformational space that must be accurately scored against computational expense. Since the entire helix is constructed as a whole from the helical library, it would seem awkward to count the helical residues the same as the non-helical ones and so helical residues are excluded when determining the number of clusters to optimize. For the loops described in this paper, this often had little consequence. For a 17-residue loop with a four-residue helix the maximum number of clusters, set at 50, is reached. The most common helical size was four residues (see Figure 4.2, below). For a 16-residue loop with a four-residue helix, the number of clusters is 48. Only for the few cases, such as PDB 2JA2, where a 16-residue loop contains an eight-residue helix, were the number of clusters, set to 32, significantly different from the maximum value of 50. These cases are the exception, and as is described later, the results from these cases, despite the reduced number of clusters, were excellent.

Side chain sampling is performed using a 10º-resolution rotamer library constructed by Xiang and Honig (107). The algorithm for side-chain optimization works by initially placing side-chains in a random rotamer state onto the backbone. Self-consistent optimization is then performed where all side-chains but one are held fixed while the free side chain is minimized. With the exception of loop prediction in a perturbed native environment, the default of one round

of side-chain randomization per entire loop minimization was found sufficient. When considering perturbed native environments, where the surrounding side chains are included in refinement, additional rounds of side-chain randomization/self-consistent optimization is performed separately to compare to predictions done without this extra sampling. The lowest energy side-chain rotamers are selected across any additional rounds of side-chain randomization. After self-consistent side chain rotamers are selected, the complete loop, with both side chains and backbone atoms, is then energy minimized. Full details about side-chain optimization are described in our past publications (108).

Scoring is done using an augmented form of the Optimized Potential for Liquid Simulations (OPLS) all-atom force field (33, 34). For solvation, an implicit model was used based on the Surface Generalized Born model as described initially in Ghosh *et al* (109). A variable dielectric approach is used to treat polarization from protein side chains (59). Additional corrections were added to the energy model to better account for π-π interactions, self-contact interactions, and hydrophobic interactions. The force field, solvation model, and all correction terms are discussed in greater detail in Li *et al.* The protonation state of all titratable residues was set using the Independent Cluster Decomposition Algorithm of Li *et al.*(110)

Since we evaluate our loop prediction method against published crystal structures, crystal-packing effects were taken into consideration. The crystallographic asymmetric unit, as well as all atoms from other surrounding unit cells that are within 30 Å, are included in the simulation. The coordinates of all copies of the asymmetric units are updated for steric clash checking and energy calculation throughout the course of the loop prediction.

Figure 4.1 Loop-helix-loop predicted in PDB 1BKR. The target loop-helix-loop residues are highlighted red from residues 75 - 91. The helix of interest, labeled α4, spans residues 82-85. Loop prediction without the helical library would assign the closure residue to be residue 83, highlighted in white. The LHL method places the closure residue at position 80. This figure was generated using ESPript.

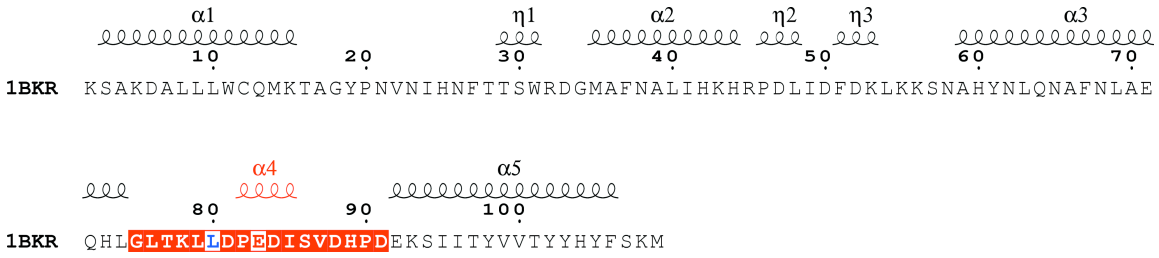
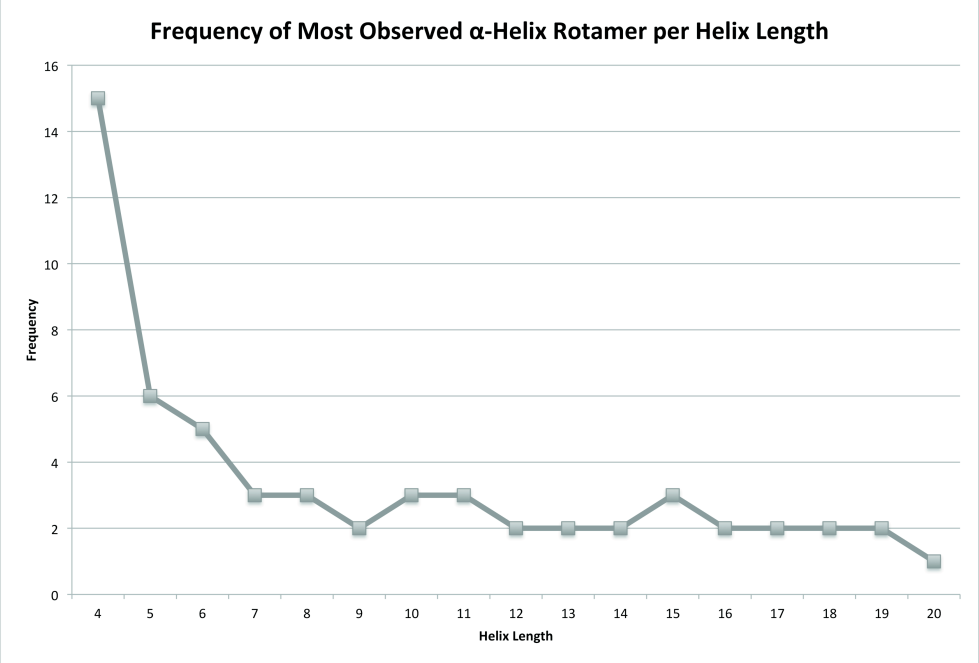


Figure 4.2 Plot of the greatest frequency observed of an α-helix rotamer per helix length. After a six residue α-helix, rotamers were only observed no more frequently than three times.



4.3.4 Construction of the helical dihedral library

As a natural extension to the dipeptide dihedral library, we constructed a helical dihedral library to exploit the coupled torsions present in an α-helix. An initial set of PDB structures was obtained from the precompiled culled PDB lists from the PISCES web server(100). The parameters used to cull the structures were a percentage identity cutoff of 30%, a resolution cutoff of 2.0 Å or better, and an R-factor cutoff of 0.25. The PDB list was obtained on October 16, 2007. The list contained 3900 PDB structures. Using an internal PLOP implementation of the DSSP algorithm (103), α-helices were identified with lengths ranging from four to twenty residues. The ϕ,ψ angles for the helical residues were extracted. We ignored values for the ω dihedral and instead used 180º during loop buildup. Deviations from the *trans* conformation are permitted during loop minimization. The dihedral angles were rounded and binned to a 10º resolution. The frequency of each binned helical rotamer was counted per helix length. In structures containing homomultimeric proteins, the helix was only counted once. We did not include helical fragments from larger helices as part of the set of dihedrals for smaller helices. That is, the torsions in a 6-residue α-helix are kept separate from the torsions in a 4-residue α-helix. This adherence to the use of only complete helices was rigidly followed throughout loop prediction. Specifically, loop buildup from both ends of the loop was done such that the helix was not divided between both loop halves. When predicting a subsection of a loop, as is done during hierarchical loop prediction, in any instance where a subsection of the helix was predicted, the dipeptide dihedral library from Zhao *et al.*(104) was used instead.

Initially, we sought to include all rotamers observed with a frequency above a set cutoff. However, this approach was problematic. Despite the large number of PDB structures, for large helices, many rotamer sets do not appear more than once. For example, in a 9-residue helix

containing 18 dihedral angles (ϕ,ψ), a single 10º difference in any ϕ,ψ angle would place that rotamer in a new bin. For helices of this length, a helical rotamer was not observed with a frequency greater than twice (Figure 4.2). Beyond a six-residue α-helix, rotamers were observed no more frequently than three times. We therefore felt that there was no suitable frequency cutoff to use. Ultimately, we arbitrarily decided to set the library to contain *2xLength_helix* rotamers and populated the library with the most frequent rotamers that conformed closest to ideal helical dihedral angles of (ϕ,ψ) = (-60°, -40°). Any non-ideality in a helix was left to be predicted during loop minimization and the multiple stages of loop refinement described in the following section.

4.3.5 Hierarchical loop prediction

Hierarchical Loop Prediction was first described by Jacobson *et al.* in 2004 and then expanded by Zhu *et al.* in 2006. In short, multiple runs of PLOP are performed where increasing constraints are applied to subsequent rounds of loop predictions. The lowest energy loops from each PLOP run are passed onto subsequent, constrained rounds of refinement. The lowest energy loop across all PLOP runs and all constraints is considered the final structure.

Hierarchical loop prediction is begun with an initial set of candidate loops that are predicted by running PLOP at discrete values of the overlap factor (*ofac*). In this work, we permitted the *ofac* to vary from 0.3 to 0.7 in increments of 0.05. The best 15 loops, in terms of energy, are passed onto a *Ref* stage. A *Ref* stage constrains the C_α atoms of any new prediction to lie within a set radius of the C_α coordinates of the previous stage. In this case, the *Ref1* stage used a 4 Å radius. The best 20 loops from this stage are passed onto a *Fix-n* stage. In a *Fix-n* stage, we repredict a subset of the original target loop but use the output from a previous stage as the scaffold, holding a total of *n* terminal residues fixed. For example, in a *Fix3* stage, we

hold three terminal residues fixed, and repredict the interior loop residues that remain. There are a total of four possible ways to fix three terminal residues:

1. Fix three N-terminal residues
2. Fix three C-terminal residues
3. Fix two N-terminal residues and one C-terminal residue
4. Fix one N-terminal residue and two C-terminal residues

All four possibilities are explored when selecting the lowest energy loop from the *Fix3* stage. In general, there are *n+1* possible combinations for a given *Fix-n* stage. We ran a total of eight *Fix* stages from *Fix1* to *Fix8*. The *Fix1* stage passed the top 10 loops onto *Fix2*. Each subsequent *Fix* stage passed one less loop onto a subsequent stage so that the *Fix8* stage passed only the top three predictions. Finally, a second *Ref* stage is run, *Ref2*, where a 6 Å C_α constraint is used. In total, taking into account all permutations in the *Fix* stages as well as the *Init* stage and *Ref* stages, there is a minimum of 334 PLOP runs per hierarchical loop prediction. The minimum number of PLOP runs can be exceeded by adaptively varying the *ofac* during hierarchical loop prediction, described in greater detail below.

To accommodate our helical dihedral library, we modified hierarchical loop prediction method in two ways: The generation of our helical library was based on complete helices. To be precise, the helical library for four-residue helices is taken only from the coordinates of helices that are exactly four residues. We do not include in our four residue helical library segments of, for example, an eight-residue helix spanning four residues in length. As such, we do not construct our loops using a separate set of "partial" secondary structural elements. As a result of this, *Fix* stages that would constrain part of a helix, instead revert to using our general dihedral library for the individual PLOP run. The use of a helical library also resulted in a large number

of individual PLOP runs that failed to produce any candidate helices. This can happen under normal circumstances, say, during a late *Fix* stage where the majority of the loop is kept constrained and only a small subset of the loop is resampled. Loop construction in these late *Fix* stages requires the residue buildup to occur without violating our *ofac* criterion despite being in an environment made all the more crowded by the unconstrained segments of the loop. This problem becomes compounded when working with a helical library. Since loop buildup with a helical library appends the helix onto a nascent loop in a single step, a slight displacement of the preceding residue leads to a large displacement of the terminal end of the helix – a sort of lever effect. If this crude displacement of the terminal residue of a helix places the loop in a steric clash with the surrounding environment, the loop candidate could be rejected due to the *ofac* criterion. In these cases, the outcome of a loop prediction becomes all the more sensitive to the *ofac* parameter. To further decouple the effect the *ofac* has on a successful loop prediction, any individual PLOP run beyond the *Init* stage that fails to succeed past loop buildup is automatically rerun with a lower *ofac* down to the lowest *ofac* sampled during the *Init* stage. In a PLOP run, the rate-limiting factor is during side chain optimization/minimization, rather than during loop buildup. Restarting a PLOP job after a failed buildup stage is on an order of magnitude of one minute. Since this procedural augmentation can apply to loop-helix-loops as much as it can to other loops, this improved sampling adjustment was applied to all cases studied in this work, regardless of the dihedral library used.

4.3.6 Calculation of RMSD

The success of loop prediction was gauged by using the backbone RMSD calculated against the native, crystal structure conformation of the loop. RMSD was calculated by superimposing the protein backbone, excluding the loop, and using the N, C_α, and C coordinates

of the loop to compute the deviations. Unless otherwise stated, we report the RMSD for the lowest energy predicted loop.

4.3.7 Calculation of the relative energy

Similar to RMSD, at the conclusion of complete hierarchical loop prediction, we report the relative energy of our predicted structure against the energy of the minimized native. This relative energy is defined as $\Delta E = E\_prediction - E\_native$. A final structure that has a poor RMSD but a calculated energy that is erroneously superior to the native would thus have a negative ΔE and would indicate a failure of our energy model. Minimization of the target for comparison against predictions is necessary to permit a fair comparison between structures but is particularly important when comparing to crystal structures as the PDB structures obtained have, in all the structures examined in this paper, no explicit hydrogen atoms. The minimization of the native was performed similarly to minimization/optimization of candidate loop structures as described above in the Single-Loop Prediction subsection of the methods. For the native, the target loop is first minimized followed by side chain sampling using the protocol described above in the Single-Loop Prediction section. For predictions done in a perturbed native environment, ΔE reports are still against the energy of the minimized native. For these cases, all additional surrounding residues that are included in the prediction are also minimized in the native to permit an accurate comparison. In instances when we used additional rounds of side chain sampling, the native loop, during minimization, was also permitted identical number of additional side chain sampling.

4.3.8 Sequence based secondary structure prediction

Loop prediction using the helical dihedral library requires the user to provide a range of loop residues, known as the helical bounds, over which to apply this library. To serve as an

initial test of our method without the complication of uncertainty in the existence and size of a helix, we predicted loop-helix-loops from previously published crystal structures. In these experiments, the helical bounds were known *a priori*. After we had observed success using exact helical bounds, we tested the robustness of this method in a more realistic setting where the helical bounds were supplied by popular sequence-based secondary structure prediction software. Specifically, we ran local copies of the secondary structure prediction packages SSPro(111) and PSIPRED (63). The output of either of these programs is a secondary structure assignment across each of the residues contained in the protein chain of interest. We examined the secondary structure assignments only for the residues that spanned our particular loops. Often times, these assignments labeled more than one set of intra-loop residues as helical. In particular, the loops discussed in this paper are sometimes bounded by larger helices and these secondary structure assignment algorithms had occasionally assigned the terminal residues of the loop to be a part of that larger flanking helix. In other cases, three, two or even a single intra-loop residue was assigned as helical. As the loop-helix-loop prediction method described in this paper is intended for α-helices (helices of four residues or larger), assigning less than four residues as helical is not useful for our purposes. Thus, for simplicity, the largest intra-loop helical segment predicted by SSPro4 or PSIPRED, spanning at least four residues, was used as the inputted helical bounds. When both PSIPRED and SSPro4 offered useable helical bounds, we performed loop prediction with both bounds separately and compared the results.

4.3.9 Loop prediction in an Inexact Environment

Unless otherwise noted, all loop predictions in this work were done by deleting the loop residues but leaving all surrounding side chains intact, thereby preserving the crystal structure environment. In an actual homology modeling experiment, the surrounding side chains are

unlikely to be placed *a priori* in their correct native conformation. To test the effectiveness of our method in refining loops in an inexact environment, we followed the approach of Sellers *et al.* (112) to perturb the surrounding side chains to a reasonable but non-native conformation. To do this, we ran multiple rounds of PLOP to predict the loop of interest in the crystal structure and selected a loop with a backbone RMSD of no better than 3 Å. A list of surrounding residues is obtained by noting all residues that are within 7.5 Å of any candidate predicted loop, not just the one loop with a 3 Å RMSD. The union of the side chains from the surrounding residue list as well as the loop side chains is minimized with the 3 Å backbone RMSD loop held in place. At this point, the surrounding side chains are "biased" towards the 3 Å RMSD loop. This structure then provides the surrounding environment for subsequent tests of our loop prediction methods.

4.3.10 Dipeptide Rotamer Frequency Score

For a number of challenging cases, we experimented with the use of a new addition to our energy model that penalizes loop conformations that are constructed with seldom-observed dipeptide dihedrals. The dipeptide rotamer frequency-based scoring term employed a greatly expanded dipeptide rotamer library (garnered from ~7500 high-quality PDB structures) that incorporated the frequency of each rotamer in this subset of the PDB. This information was used to penalize loop dipeptides whose combination of (ϕ,ψ) angles fall in an extremely unpopulated region of the five-dimensional dipeptide analogue to the well-known Ramachandran plot. The set of five angles for each dipeptide in the predicted loop, using a "sliding window" scheme, is compared against the new library to find the nearest dipeptide rotamer. Two criteria determine whether a penalty will be applied to the dipeptide:

1. If the Euclidean distance between the loop dipeptide and the nearest rotamer in the library is greater than a certain, empirically determined cutoff.

2. If the total population of rotamers within a set radius of the loop dipeptide is below a certain threshold.

The form of this penalty term, its implementation, and its successes in improving loop prediction in crystal structure and homology model environments will be discussed in detail in an upcoming publication. This term was used in two situations:

1. For all of the predictions in inexact environments. This is a substantially more challenging sampling and scoring problem, and the information contained in the dipeptide score can be expected to improve results systematically.
2. For a small subset of the predictions in the native environment where difficulties in the standard prediction approach were encountered.

To date, we have not found any cases where this term worsens results. However, more extensive tests are underway and will be presented in a subsequent publication

4.4 Results and discussion

4.4.1 Description of test cases

Application of the discriminating criteria used to select suitable LHL test cases yielded a set of 35 loop-helix-loops of 16 or 17 residues in length. These loops exhibited a distribution of helix size as shown in Figure 4.3. The distribution indicates a diversity of helix sizes within a 16- or 17-residue loop. Although the helical library described in this work is only for α-helices, loops were included that contained 3^{10} helices, either separate from an α-helix already present in the loop, or as the sole secondary structure of the loop. It is these former cases where a loop contains both a 3^{10} helix with an α-helix that led to the non-zero frequency for helices of length three (Figure 4.3).

PDB 1W27 contains a noteworthy example of a multi-helical loop. The 17-residue loop,

contains a 4-residue 3^{10} helix and 5-residue α-helix separated by a single residue, D302 (Figure 4.4). Evidently, residue D302 permits flexibility in the backbone to transition from one helical type to another. We explored the use of our α-helical library in three approaches: 1) Loop prediction given the α-helix as the helical bounds; 2) Loop prediction given the 3^{10}-helix as the helical bounds; 3) Loop prediction where the 3^{10} and α-helix bounds are combined to yield a 10-residue "α-helix." The results of these approaches are described in greater detail below.

PDB 2VPN was another case of a multi-helical loop. The 16-residue loop of interest is composed of a 4-residue α-helix and a 7-residue α-helix separated by a single residue, E102 (Figure 4.5). Residue E102 is kinked, according to DSSP, failing to form the periodic hydrogen bond expected of an α-helix. As in the 1W27 case, we tried three approaches to predicting this loop. For β-hairpins, a set of 41 cases was collected satisfying the criteria described in the methods section. The size of the hairpin region ranged from 6 to 13 residues within loops up to 17 residues in length. Hairpin size is defined to be the number of residues from the start of the first β-strand to the end of the second β-strand, including all non-β residues in between. Hairpins occurred most frequently as either six or eight residues in length (Figure 4.3). However, since the formation of the coordinated hydrogen bonds is what is most challenging in loop-hairpin-loop prediction, we feel it is useful to describe the distribution of hydrogen bonds across our set of β-hairpins. Hairpins contained from four to eight hydrogen bonded residues with the number of coil/turn residues contained within the hairpin ranging from two to seven residues (Figure 4.6). Thus, this test set of β-hairpin containing loops required the successful prediction of at least one specific hydrogen bond spanning at most seven residues.

Figure 4.3 Distribution of secondary-structural elements within the test set of loops. Helices of length 3 were from 3^{10} helices found in loops already containing an α-helix. Hairpin length includes the terminal hydrogen bonded residues as well as all residues in between.

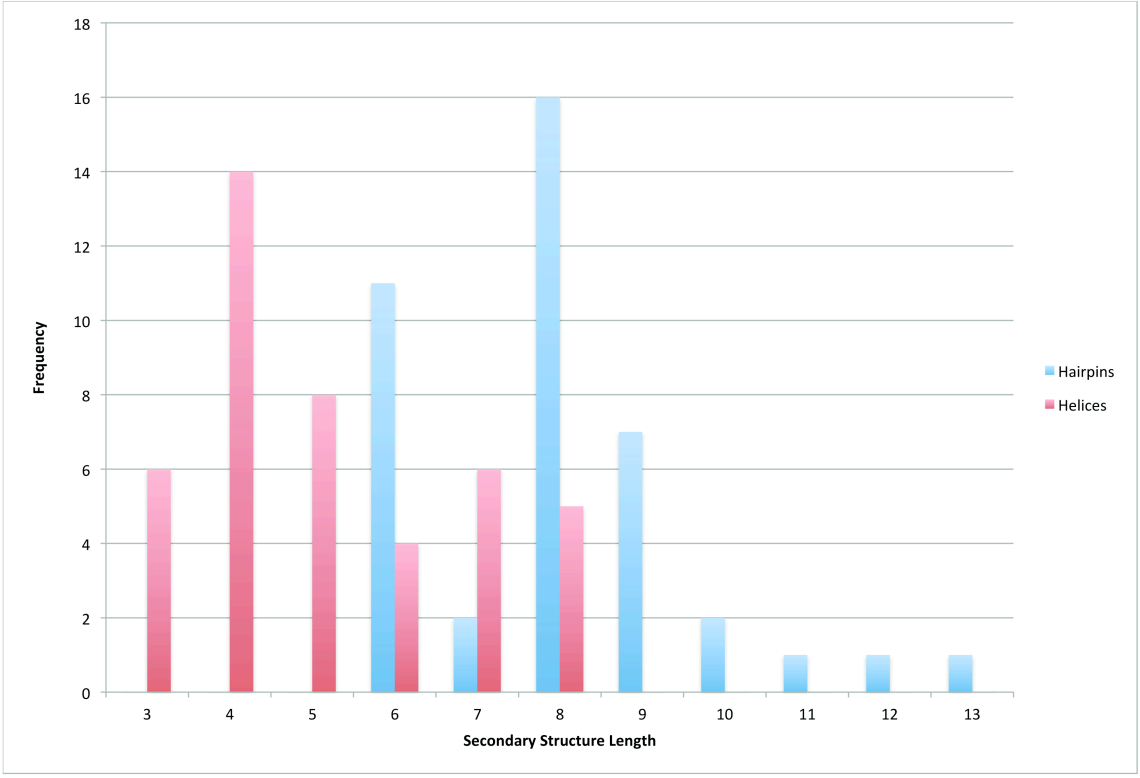
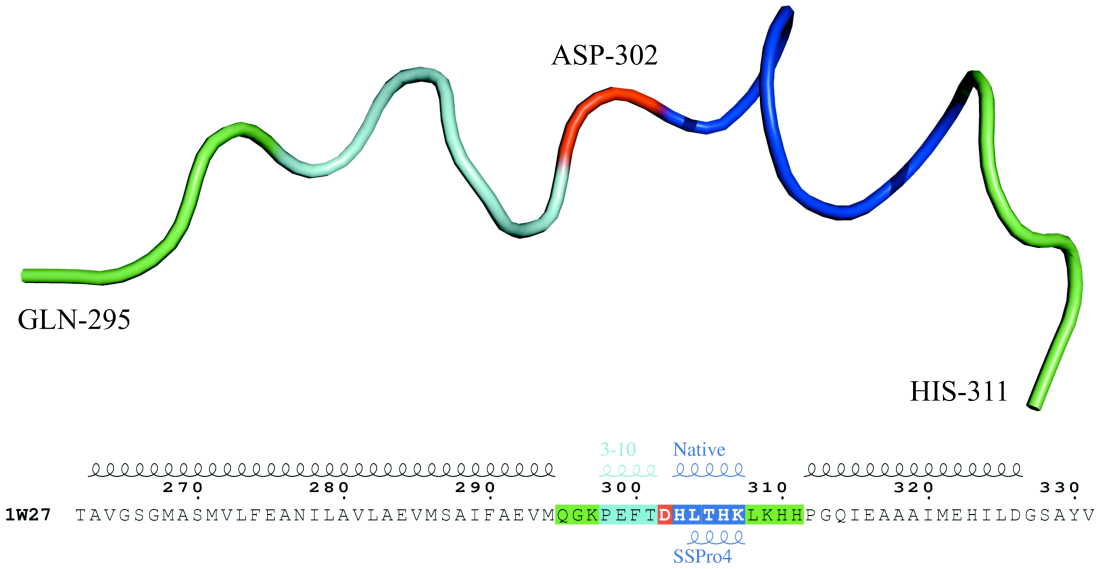


Figure 4.4 Multi-helical loop in PDB 1W27. The loop bounds are Q295 to H311. Residues preceding and following the helices are colored green. The 5-residue α-helix is colored blue while the 4-residue 3^{10}-helix is colored cyan. Residue D302, the kinked residue dividing the two



helices, is colored red. We attempted separately to use the helical bounds of either the α-helix, 3^{10}-helix, or treated all ten residues as one "α-helix". SSPro4, a sequence-based secondary structure prediction program, assigned the four residues from L304-K307 as helical. The sequence annotation was generated using ESPript (113). This loop confirmation, and all other similar illustrations were produced using Pymol (114).

Figure 4.5 Multi-helical loop in PDB 2VPN. The loop bounds are S97 to G112. Residues preceding and following the helices are colored green. The 7-residue α-helix is colored cyan while the 4-residue α-helix colored blue. Residue E102, the kinked residue dividing the two helices, is colored red. We attempted separately to use the helical bounds of either the seven-residue helix, the four-residue helix, or treated all twelve residues as one "α-helix".

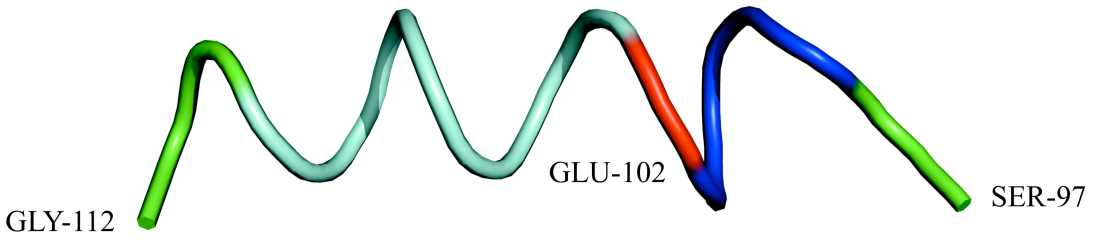
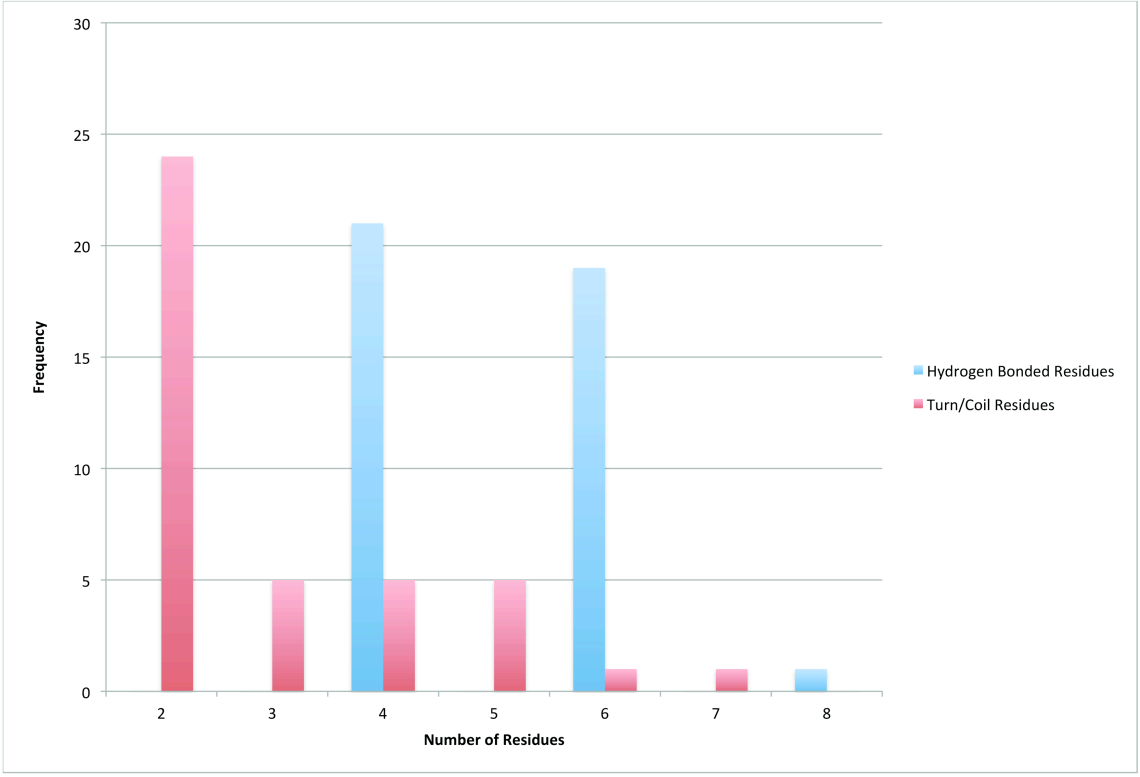


Figure 4.6 Distribution of hairpin characteristics. Hairpins contained from four to eight hydrogen bonded residues and with the internal turn/coil residues spanning a length from two to seven residues.



4.4.2 Predictions performed in the crystal structure environment

A total of 35 loop-helix-loop (LHL) cases and 41 beta-hairpin cases were predicted in the crystal structure environment. In the crystal structure environment, the loop of interest is deleted and rebuilt while the surrounding residues remain fixed. In this work, we compare the predictions done using a helical dihedral library versus predictions performed using the standard PLOP dihedral library (104).

4.4.3 Loop-helix-loops predicted using the dipeptide dihedral library versus the helical dihedral library with exact helical bounds

As a first test of the helical dihedral library, we performed loop prediction on the set of 35 LHL cases either with the previous dipeptide dihedral library or with the helical library described in this work. Experiments such as these were primarily meant to ensure that in the absence of uncertainty in the size and location of the helix, our helical library method could succeed. A prediction performed where the helix is postulated from secondary structure prediction software is our primary methodological algorithm to be used in realistic prediction situations, and is discussed later. Table 4.1 provides a summary of the results as a function of helix length. Compared to the dipeptide dihedral library, the helical dihedral library consistently displays improved accuracy, with mean and median RMSD always below 1 Å. No strong correlation is noted between the size of the internal helix and the results from either dihedral library. This suggests, consistent with past results, that the difficulty in loop prediction lies with the size of the loop, rather than the secondary structure contained in the loop, at least for helices up to eight residues in length.

For LHLs containing a four-residue helix, both dihedral libraries appear to perform similarly. As might be expected, the helical library shows the greatest advantage for predictions

containing an eight-residue helix with superior median and mean RMSD values by around 0.5 Å. It is likely that the coordinated hydrogen bonds that need to be formed are easily generated when explicit helical dihedrals spanning the precise residues are deliberately introduced during sampling. This seems particularly relevant for the LHL in PDB 2YR5. This is a 16-residue loop containing a 7-residue α-helix (Figure 4.7).

The dipeptide dihedral library produces a 7.26 Å RMSD loop with a ΔE of -0.9 kcal/mol relative to the minimized crystal structure, while the helical dihedral library leads to a 1.11 Å RMSD loop with a ΔE of -18.34 kcal/mol. The dipeptide dihedral library clearly fails to form the native helix, forming instead a loop that protrudes out in solution. The prediction with the helical library is dramatically superior but forms a larger nine-residue α-helix. Evidently, the shorter seven-residue α-helix "seeds" the larger helix. Considering the large negative ΔE energy relative to the native, these additional two helical residues may be the result of an energy error incorrectly favoring formation of additional helical residues. While slightly detrimental to the accuracy of this particular loop prediction, as is discussed in greater detail below, the use of a shorter helix to "seed" a larger one is later exploited to find the lowest energy loop.

Two PDB structures, 1W27 and 2VPN, each contain a multi-helical loop-helix-loop that still satisfied the criteria stated above for selecting loops (Figure 4.4, Figure 4.5). These cases provided an opportunity to explore the effect of the helical dihedral library in complex situations. We attempted to predict the loop by supplying as helical bounds either of the two helices or treated the helices as combined, disregarding the non-helical residues dividing the helices. Table 4.2 describes the result of these loop predictions. In both cases, the helical library produced the lowest energy conformation with sub-Ångström RMSD.

Figure 4.7 Loop-Helix-Loop predicted in PDB 2YR5. The native loop coordinates are colored blue with the 7-residue α-helix colored teal. The prediction using the helical dihedral library is shown in red with the resulting 9-residue α-helix colored in pink. The loop prediction performed using the dipeptide dihedral library is shown in green. Despite supplying the exact 7-residue helical bounds during loop prediction with the helical library, what resulted was a slightly larger helix, evidently "seeded" by the smaller 7-residue α-helix.

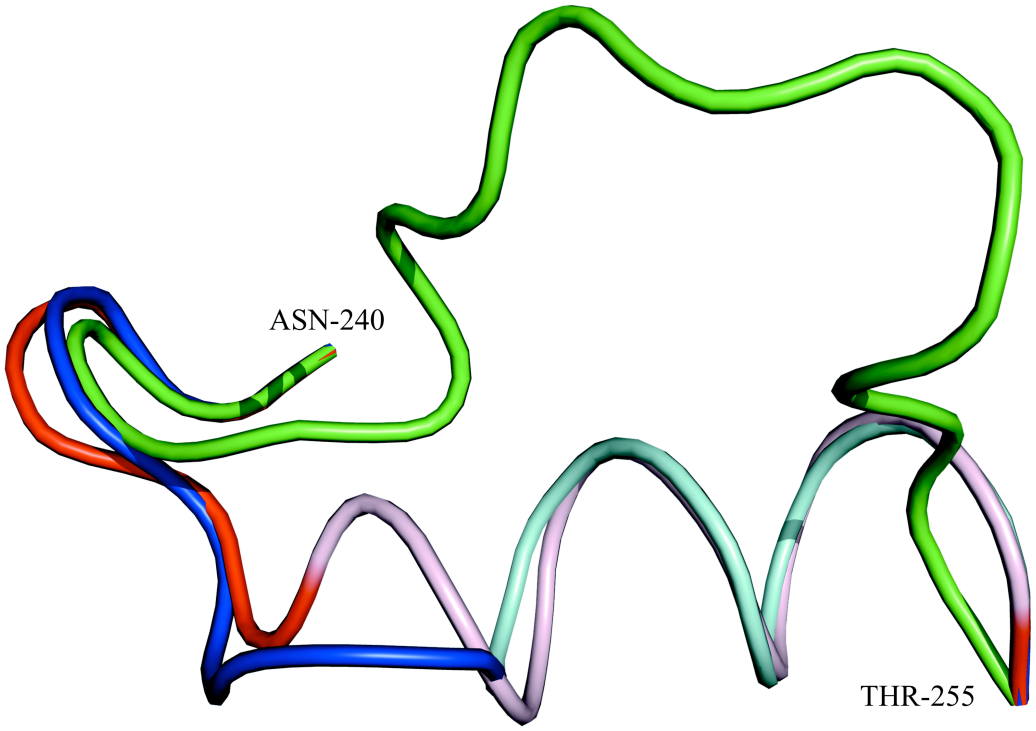


Table 4.1 Comparison of Loop-Helix-Loop predictions with the dipeptide dihedral library versus the helical dihedral library. The two noteworthy multi-helical loops found in PDB 1W27 and 2VPN are excluded in this table. The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction and corresponds with the ΔE.

Helix Length	Number of Cases	Dipeptide Dihedral Library				Helical Dihedral Library with Exact Helical Bounds			
		RMSD (Å)		ΔE (kcal/mol)		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean	Median	Mean	Median	Mean
4	12	0.53	1.29	3.89	12.39	0.55	0.99	-0.02	-3.18
5	7	0.91	1.09	-7.94	-6.19	0.51	0.80	-3.74	-3.41
6	4	1.00	0.95	2.06	11.11	0.62	0.77	2.47	2.75
7	5	0.55	1.94	0.51	5.19	0.79	0.91	2.52	1.59
8	5	0.81	0.98	4.33	6.66	0.36	0.41	-4.22	-2.18

Table 4.2 Prediction of multi-helical loops using various loop bounds. When no helical bounds were supplied, loop prediction was performed using the dipeptide dihedral library. The 1W27 prediction using the 4-res 3^{10}-helix for helical bounds still employed the α-helix dihedral library described in this work. The combined helical bounds of 1W27 and 2VPN consider both helices to be one large α-helix during loop buildup. The truncated SSPro helix is equivalent to the 5-res α-helix but truncated one residue at the helical N-terminus. ΔE refers to the change in energy of the predicted loop relative to the native conformation.

PDB	1W27					2VPN			
Helical Bounds Supplied	None	4-res 3^{10}-helix	5-res α-helix	Combined 10-res "helix"	SSPro truncated α-helix	None	4-res α-helix	7-res α-helix	Combined 12-res "helix"
RMSD (Å)	2.69	1.50	0.77	1.98	0.34	0.42	0.41	0.37	0.38
ΔE (kcal/mol)	38.96	22.27	-3.43	24.32	-12.19	2.01	11.08	-9.69	0.23

4.4.4 Loop-helix-loop prediction based on helical bounds derived from SSPro4 and PSIPRED

In the previous section, exact helical bounds were used which were taken from the output of DSSP when applied to the crystal structure. Such accurate information will not be known *a priori*. Indeed, significant variability in the definition of secondary structure assignment has been known to affect the precise bounds of secondary structure, especially as the number of secondary structure assignment definitions is now legion. To simulate the effectiveness of using the helical dihedral library in more realistic computational experiments, and to further gauge the sensitivity of our method to accurate knowledge of the helical bounds, we applied the popular sequence-based secondary structure prediction packages SSPro4 and PSIPRED to our set of 35 loop-helix-loops and attempted loop prediction using these predicted helical bounds. The results from these secondary structure prediction packages, excluding the multi-helical loops of PDB 1W27 and 2VPN, are presented in Table 4.3.

Comparing the two packages, it would appear that SSPro4 could more reliably find exact or overlapping helical bounds compared to PSIPRED, however the two methods are complementary. For example, SSPro4 fails to find any helix in the LHL in PDB 3LY0, while PSIPRED found a truncated helix whose bounds are contained within the DSSP results. We must caution the reader that we do not attempt here to perform a rigorous evaluation of secondary structure prediction algorithms. For that, we refer the reader to Koh *et al.* 2003 (115) and Pirovano and Heringa, 2010 (116). Rather, we simply selected two popular and easily available packages for our study. Alternative secondary structure prediction algorithms may be just as valid, as is using more than two packages to find the helical bounds. However, the fact that in a large set of cases, the exact, DSSP helical bounds were identified provides some legitimacy in

interpreting the results from the previous section – accurate knowledge of a helix within an LHL is not unreasonable.

For the two multi-helical loops in PDB 1W27 and 2VPN, the two secondary structure prediction methods contrast. For the LHL in PDB 1W27 (Figure 4.4), PSIPRED correctly identifies the five-residue α-helix but fails to predict the four-residue 3^{10}-helix. SSPro4 also fails to identify the 3^{10}-helix but the α-helix is incorrectly predicted to be four residues, truncated at the N-terminus. In 2VPN (Figure 4.5), PSIPRED predicts a combined helix that spans both α-helices and extends one residue further towards the C-terminus. Contrastingly, SSPro4 considers the entire LHL to be one large helix – a result that is inadequate for our helical dihedral library approach. In both of these cases, PSIPRED offers a reasonable set of helical bounds for use in our method.

Table IV summarizes the results of LHL prediction using the helical bounds, when available, from PSIPRED and SSPro4. In general, the helical bounds provided by the sequence-based secondary structure prediction methods SSPro4 and PSIPRED are effective in loop-helix-loop prediction. Although the statistics might suggest that the fewer cases afforded by PSIPRED result in higher quality predictions, we refrain from making such a conclusion, as it may be necessary to also take into account the size of exact helix studied. This does illustrate, however, that sequence based secondary structure assignments are useful to our method when performing three-dimensional loop prediction.

It should be mentioned that five cases were found where the helical bounds offered by either PSIPRED or SSPro4 resulted in failed loop predictions where not a single predicted loop was constructed. In four of these five cases (PSIPRED bounds: PDBs 1N45, 1OAO, 2YR5; SSPro4 bounds: PDB 3GWI), the sequence-based secondary structure assignment places the

helix as part of the N or C terminus. It would appear that in these cases, the sequence-based assignment is extending the larger helix that forms the boundary of the loop-helix-loop into what DSSP, and the criteria used in this paper, consider to be part of the loop. Although in practice, assigning the terminal residues of a loop to be helical is not fatal – PSIPRED and SSPro4 both place a helix on the C-terminus of the LHL in PDB 1HN0 and yet a sub-0.5 Å RMSD loop is predicted – loop prediction without any non-helical residues to precede the helix is extremely difficult.

In these situations, the lever effect, described previously in the single-loop prediction section of Materials and Methods, becomes very pronounced. As PLOP constructs the loop in a tree-based method, where the tree is split into additional branches as more loop residues are predicted, placing the helix at a loop terminus means there are no preceding branches to rely upon. Whatever few positions the leading residue of the helix is placed at are set entirely by the sparse number of helical rotamers present in our library. In practice, this means that all the rotamers in our helical rotamer library for a given helix size are easily rejected. Although in principle, one could reduce the *ofac* parameter to permit greater steric overlap between a loop residue and the surrounding environment, in practice, the *ofac* was rarely seen as the limiting factor. The one case that permitted loop prediction after adjusting the *ofac* was the PSIPRED bounds for 1N45, however, we had to set the *ofac* to an abnormally low value of 0.20, meaning enormous steric clashes were permitted. Even still, the output of this loop prediction only produced a 5.69 Å RMSD loop with a ΔE of 9.30 kcal/mol.

In all cases, nascent loop segments were screened out when the helix placed a residue too far from the body of the protein to what has been empirically observed across published crystal structures containing protein loops. Or instead, loops were screened when the distance between

the loop segment containing the helix and the opposing end of the loop is considered too great to be spanned by whatever intermediate residues remain. In other words, the helix places one half of the loop too far away for loop closure to be possible. These loop screening methods are described briefly in the Materials and Methods section, and in greater detail in Jacobson *et al*. Setting the *ofac* to an arbitrary low value has no effect on these screens – the helical rotamer library simply does not contain a suitable rotamer to permit loop prediction with the supplied helical bounds. Although there is certainly an argument to be made for increasing the size of the helical library, as evidenced from our other successes, the size of the library does not appear to be an impediment to loop-helix-loop prediction. Rather, the practitioner of our method might gain insight by noting that if no suitable rotamer is present in the library, it may be prudent to consider alternative helical bounds. Indeed, none of these terminus-bounded helices are the crystal structure helical bounds – we avoided such cases by our definition of loop-helix-loops. Determining the helical bounds from the output of our previous dipeptide-dihedral library method, as discussed in greater detail below, may be a fruitful alternative. The multi-helical loop of 2VPN (Figure 4.5) is one slight exception. In this case, PSIPRED combines the 4-residue α-helix and the adjacent 7-residue α-helix into one large helix and even extends the helical bounds further by one additional residue to produce a 13-residue helix. SSPro4 simply considers the entire loop-helix-loop to be one large helix, an outcome useless for our helical dihedral library. In this case, the helical bounds provided by PSIPRED produce independent N- and C-terminus loop segments but closure is not achieved. This result occurs regardless of how low we set the *ofac*. Again, extending the size of the helical library may offer a solution to this case, but more likely, the helical bounds provided deviate too greatly from the native structure to permit reasonable loop prediction.

Table 4.3 Results of sequence-based secondary structure prediction packages PSIPRED and SSPro4 on our set of LHLs, excluding cases 1W27 and 2VPN, the multi-helical loops. Exact helical bounds are those that are in perfect agreement with the bounds assigned by DSSP on the crystal structure. Truncated helical bounds are those that lie within the DSSP assigned bounds. Helical bounds are considered overlapping if the secondary structure predicted helix has at least a single residue overlapping the exact bounds. No helix is considered predicted if the entire loop-helix-loop lacks any helical assignments greater than three residues.

Helical Bounds Predicted	PSIPRED	SSPro4
Exact	2	14
Truncated	6	2
Overlapping	6	9
Non-overlapping	1	1
No Helix	18	7
Total	33	33

Table 4.4 LHL prediction using the helical bounds available from PSIPRED and SSPro4. Multi-helical cases 1W27 and 2VPN are included in these statistics. Cases where the helical bounds provided by sequence-based secondary-structure prediction are not useable in our method are excluded. Further, cases where no loops were able to be predicted with the supplied helical bounds are also excluded.

Method	Number of Successful Cases	RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean
PSIPRED	13	0.44	0.49	-1.37	-1.54
SSPro4	25	0.60	0.91	1.05	0.65

4.4.5 Truncated helical bounds from sequence-based secondary structure prediction or derived from inspection of coordinates predicted with the standard PLOP dihedral library

In a few cases, sequence-based secondary structure prediction methods produced a helix that was truncated relative to the native helical bounds, yet these cases performed as well, if not better, than the native bounds. For example, PDB 1W27, one of the multi-helical loops, is composed of a four-residue 3^{10}-helix and an adjacent five-residue α-helix (Figure 4.4). SSPro4 fails to identify the 3^{10}-helix but predicts the α-helix to be truncated by one residue at the helical N-terminus, relative to the exact helical bounds (Figure 4.4). PLOP was able to predict this LHL with an RMSD of 0.77 Å and a ΔE of -3.43 kcal/mol when using the native, five-residue α-helix. However, the SSPro4 bounds led to a predicted LHL with a superior RMSD of 0.34 Å and a ΔE of -12.19 kcal/mol. Table 4.2 summarizes these results. These loop predictions are illustrated in Figure 4.8.

Consistent with our past discussion, the smaller helix may permit less of a lever effect and thereby enable finer sampling of the α-helix. It should be noted, however, that the absence of any helical bounds, that is, using the previous dipeptide dihedral library from our previous work, results in a 2.69 Å RMSD prediction (Table II). Thus the small helix is shown to also seed our hierarchical sampling method to more heavily explore conformational space near α-helices.

The LHL in PDB 2YR5 is another case where truncated helical bounds led to a superior prediction. However, this is one of the cases where the helical bounds provided by both PSIPRED and SSPro4 were attached to the LHL C-terminus and no loops emerged from our attempts at predicting this LHL with such helical bounds. Rather, we attempted LHL prediction using as helical bounds all possible four-residue α-helices that lie within the 10 residue α-helix suggested by PSIPRED and SSPro4 – a set of seven possible helical bounds. Both PSIPRED and

SSPro4 suggested identical helical bounds. The results from these predictions are shown in Table 4.5.

The predictions indicate that nearly every possible four-residue α-helix attempt produces results that are nearly identical to the LHL prediction performed using the native, seven-residue α-helix. While knowledge of the precise, native helical bounds may not be available, we demonstrate that we can still exploit information provided by sequence-based secondary structure prediction, even if that information does not perfectly match the DSSP secondary structure identification obtained from the crystal structure of the native conformation.

In total, we attempted all possible four-residue α-helix bounds for all LHL cases where the lowest energy loop was found only by using the native helical bounds. This was performed in order to discount the concern that precise *a priori* information about a helix must be known. In many cases, information about a helix was provided by sequence-based secondary-structure prediction. However, as we show in Table 4.1, providing no helical bounds and using the dipeptide dihedral library can still lead to low RMSD predictions and the formation of a helix. From these cases where a helix four-residues or larger was produced *ab initio*, we also applied our truncation sampling method across the predicted helix and took the lowest energy loop. When the dipeptide-dihedral library simply produced a four-residue helix, we reattempted loop prediction using the helical dihedral library with this previously found 4-residue helix as bounds. The lowest energy loops predicted from these experiments are shown in Table 4.6. In general, the truncation method produces helices that, on their own, are quite accurate with sub-Ångström RMSD routinely reported.

Figure 4.8 Loop-helix-loop prediction for the multi-helical loop in PDB 1W27. The native loop is shown in red. Loop prediction using the exact five-residue α-helix is shown in green. Loop prediction using the truncated, four-residue α-helix provided by SSPro4 is shown in blue. Loop prediction using the truncated four-residue α-helical bounds appears to permit improved sampling of the alpha helix. Notice that the greatest discrepancy between the two loop predictions occurs along the α-helix near the C-terminus.

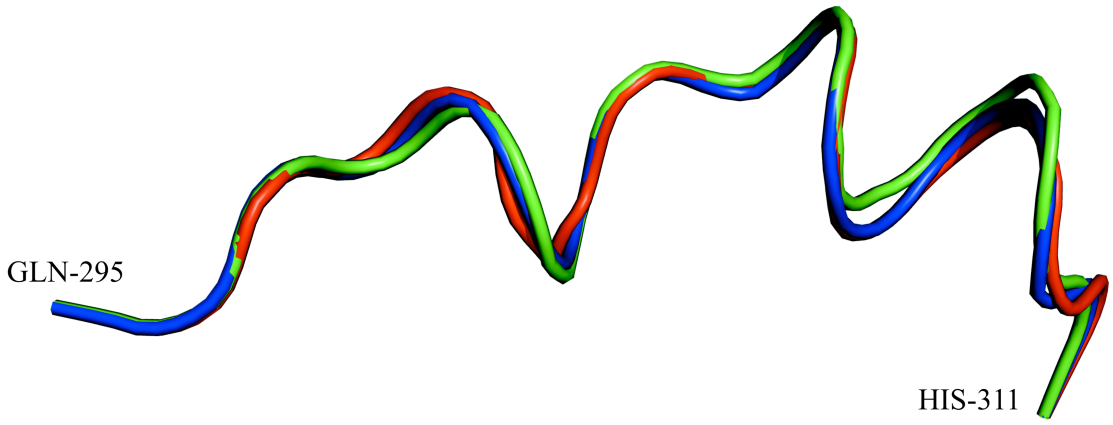


Table 4.5 Prediction results from the LHL in PDB 2YR5. LHL prediction without helical bounds refers to the use of the dipeptide dihedral library exclusively. The native bounds are those provided by DSSP analysis on the crystal structure. The PSIPRED/SSPro helical bounds are from B:246 and B:255 and bracket the seven truncation attempts shown. The lowest energy prediction across all helical bounds is highlighted in red.

Helical Bounds	RMSD (Å)	ΔE (kcal/mol)
None	7.26	-0.9
B:248 – B:254 (Native bounds)	1.11	-18.34
Bounds derived from PSIPRED/SSPro4 Truncation		
B:246 – B:249	1.11	-6.2
B:247 – B:250	1.11	-18.72
B:248 – B:251	1.11	-18.49
B:249 – B:252	1.11	-18.43
B:250 – B:253	1.10	-18.18
B:251 – B:254	1.10	-18.28
B:252 – B:255	4.27	28.57

Table 4.6 Result of LHL prediction using truncated helical bounds. All possible 4-residue helical bounds that lie within bounds provided by sequence-based secondary structure prediction or by analyzing the results from the dipeptide-dihedral based predictions were used. What is shown is the lowest energy prediction across all helical bounds attempted.

PDB	RMSD (Å)	ΔE (kcal/mol)
1HN0	0.31	-2.77
1Q1R	0.30	-8.07
1WOV	0.95	-6.08
2EX0	1.74	0.91
2FHF	0.62	-8.02
2II2	0.35	-3.4
2J9O	1.55	2.65
2QMC	0.49	-3.05
2VPN	0.22	-11.94
2YR5	1.11	-18.72
3GWI	0.53	3.77
Mean	0.80	-4.28
Median	0.58	-3.23

4.4.6 Creation of a systematic method for predicting loop-helix-loop regions

We have described above a number of different approaches to predicting LHL regions, each of which exhibits significant success for a subset of test cases. We briefly enumerate these methods below:

1. Normal loop prediction, without any use of the helical rotamer library.
2. Use of the rotamer library with helical bounds specified by the results of either SSPro or PSIPRED secondary structure prediction (this leads to two separate calculations).
3. Reprediction of the loop subsequent to normal loop prediction, using as helical bounds helical regions forming spontaneously in the normal loop prediction simulation.
4. Truncated helix loop prediction where all possible four-residue helixes that can fit within previously obtained helical bounds are explored.

Our final algorithm is a composite method in which all of the above calculations are performed for each loop, and the lowest energy prediction is selected as the predicted result. The computational cost of this composite method is roughly 4X that of one normal loop prediction. In return, one achieves a remarkably high level of reliability as is shown in Table 4.7 below. The vast majority of predictions are sub-Angström, an exceptionally low level of error for loops of this length and complexity. Only one loop has an RMSD greater than 2A, the loop in PDB 2O70. We discuss this case further below, but in essence neither normal loop prediction, nor any of the secondary structure prediction methods, predict a helix in the relevant region. When the native helix is seeded into the calculation, a superior prediction is returned. Thus, this is a sampling problem, which we can hope to solve by improving the sampling algorithm. However, with the current approach, such sampling errors are very infrequent.

Arguably, the results from predictions with the native helical bounds rely on information

that may not be precisely known in a homology modeling experiment. As such, we also report in Table 4.7 the RMSD of the lowest energy loop prediction across all sampling methods. For comparison, results of LHL prediction using helical bounds taken only from the native PDB are shown in the right half of Table 4.7.

Overall, by exploring helical bounds provided by sequence-based secondary-structure prediction methods, as well as using the truncation method, we were able to predict LHLs with slightly superior accuracy than if we were to rely on the DSSP identified helical bounds. However, there were four cases where we were unable to produce a prediction that was superior to approach using the DSSP-based bounds. Three of the four predictions are 0.11 Å from the DSSP results and can be left as acceptable.

The only egregiously inferior prediction was for the LHL in PDB 2O70. Here, the use of the DSSP-based helical bounds led to a 1.71 Å RMSD prediction compared to a 3.24 Å RMSD prediction performed solely using the dipeptide dihedral library – that is, without any supplied helical bounds (Table VII). Evidently, this LHL is a challenge for sequence-based secondary-structure prediction as well since neither PSIPRED nor SSPro4 predict there being any helix at all within the LHL. Cendron *et al*., 2007 argue that the sequence of PDB 2O70, an OHCU decarboxylase from *Danio rerio* (zebrafish), lacks homology with other known amino acid sequences (117). This may have been the case in early 2007 but evidently is now longer so. In June 2007, the crystal structure of *Arabidopsis thaliana* OHCU decarboxylase was published (PDB: 2Q37), and in 2010, the *Klebsiella pneumoniae* structure (PDB: 3O7I) was deposited in the PDB (117, 118). However, in these two more recent structures, the five residues comprising the α-helix are not conserved and the more homologous eukaryotic 2Q37 structure fails to form a helix at this position. It seems reasonable then that PSIPRED and SSPro4 would fail to identify

this helix.

With respect to size of our helical dihedral library, the LHL in PDB 1O7E posed the only challenge. In the above Table 4.7, we report the prediction results when using an augmented helical dihedral library containing the native dihedrals for the helix. In the absence of this addition to our library, the LHL prediction led to a sampling error with an RMSD of 2.09 Å and a 16.99 kcal/mol ΔE compared to a 0.37 Å RMSD and 3.34 kcal/mol ΔE with the augmented library. As discussed in the methods section, our helical dihedral library is populated with rotamers that conform close to ideality. This approach fails here and seems likely due to the large discrepancy from ideal (ϕ,ψ) angles for the two terminal residues of the helix. While we expect angles near (ϕ,ψ) = (-60°,40°), the torsions for two of the N-terminus residues of the helix, A223 and G224, are (ϕ_{A223},ψ_{A223}) = (-68°,-20°) and (ϕ_{G223},ψ_{G223}) = (-104°,1°). In particular, the terminal glycine residue poses the largest problem. From this limited case, there may indeed be utility in further expanding our helical dihedral library, but even in its current implementation, the difficulty in this LHL case appears anecdotal.

The ability of the energy model to robustly pick out the correct loop as being lowest in energy provides new confirmation of the quality of our latest generation model, supporting the results obtained in Li *et al.,* for long loop regions without secondary structure elements embedded. It is true that phase space available to the loop is significantly restricted when the native environment is (as here) retained; nevertheless, previous results from our group and others show that it is quite easy to generate grossly incorrect predictions (with substantial energy errors) with an inferior scoring function. The results discussed below in which surrounding side chains are allowed to move, in which sub-Ångström results are uniformly obtained, provides further evidence of scoring function accuracy and robustness.

Table 4.7 Results of all LHL predictions independent of helical bounds derived from analysis of the crystal structure as well as the results using bounds derived exclusively from the crystal structure. By sampling with alternate helical bounds derived from sequence-based secondary-structure prediction and/or the truncation method, the LHL.

PDB	Helical Bounds Identified Without DSSP			Exclusively DSSP Identified Helical Bounds	
	Method	RMSD (Å)	ΔE (kcal/mol)	RMSD (Å)	ΔE (kcal/mol)
1BKR	SSPro4	0.55	1.05	0.55	1.05
1E3D	SSPro4	0.55	-1.1	0.55	-1.10
1HN0	PSIPRED	0.35	-5.51	0.3	0.22
1L5W	SSPro4	0.4	-3.74	0.4	-3.74
1LLF	PSIPRED	0.44	-5.53	0.45	-5.5
1N45	Dipeptide	0.36	-4.22	2.05	12.55
1N7O	SSPro4	0.4	-1.18	0.4	-1.18
1O7E*	SSPro4	0.37	3.34	0.37	3.34
1OAO	SSPro4	0.49	-80.01	0.49	-80.01
1OX0	Dipeptide	1.35	-13.23	0.58	-8.7
1Q1R	Truncate	0.3	-8.07	0.3	-8.07
1QMY	SSPro4	1.27	-2.67	1.27	-2.67
1SU8	Dipeptide	0.43	2.24	1.45	18.65
1W27	SSPro4	0.34	-12.19	0.77	-3.43
1WOV	Truncate	0.95	-6.03	0.67	-2.88
1ZX0	Dipeptide	1.04	11.2	1.81	20.12

2DEB	**Dipeptide**	1.35	-10.14	1.55	8.93
2EX0	**PSIPRED**	0.44	-0.86	0.33	-4.4
2FHF	**Dipeptide**	0.54	-8.7	0.54	-10.16
2II2	**Truncate**	0.35	-3.4	0.36	-4.22
2J9O	**Truncate**	1.55	2.65	1.55	2.65
2JA2	**Dipeptide**	0.81	0.13	0.72	1.15
2JDI	**SSPro4**	0.51	15.02	0.51	15.02
2O70	**Dipeptide**	3.24	-12.42	1.71	-15.09
2P0W	**Dipeptide**	0.47	-7.94	0.51	-6.96
2QMC	**Truncate**	0.49	-3.05	0.49	-3.05
2RJ2	**PSIPRED**	0.31	-1.37	0.57	4.72
2V36	**Dipeptide**	0.18	-1.22	0.25	-0.37
2VPN	**Truncate**	0.22	-11.94	0.37	-9.69
2WEU	**Dipeptide**	0.91	-10.24	1.73	12.19
2YR5	**Truncate**	1.11	-18.72	1.11	-18.34
3CWW	**SSPro4**	0.28	-12.04	0.27	-10.71
3GWI	**Truncate**	0.53	3.77	0.38	7.28
3HL0	**PSIPRED**	0.93	-2.04	0.39	2.52
3LY0	**PSIPRED**	0.54	1.28	0.79	9.14
Mean		0.70	-5.91	0.76	-2.31
Median		0.50	-3.57	0.55	-1.74

4.4.7 Hairpins predicted using the standard PLOP dihedral library

In addition to loop-helix-loops, we also attempted prediction of, what could be termed, loop-hairpin-loops as another challenge of loop prediction containing local secondary-structure. The results from loop-hairpin-loop prediction, arranged by hairpin length, are shown in Table 4.8, and the complete results for all 41 hairpin predictions are provided in Table 4.9.

Similar to the results for loop-helix-loop predictions, we observe no correlation between the size of the hairpin and the RMSD of the predicted loop-hairpin-loop. We note however that one of the eight-residue hairpin cases produced a large discrepancy between the median and the median (Table VIII). This case is part of PDB 2ZBX and led to an RMSD of 17.29 Å with a surprising ΔE of -177.74 kcal/mol. It should be noted that the second best case has an acceptable RMSD of 1.02 Å and a ΔE of -10.91 kcal/mol. Of course, we cannot choose this 1.02 Å loop as the best case *a priori* as determination of the best loop is made purely on energetic grounds. The apparent lowest-energy loop and the native are shown in Figure 4.9.

However, it was observed that the dihedrals in the predicted loop occupy regions of dipeptide-dihedral space (ϕ_1,ψ_1,ω,ϕ_2,ψ_2) that are poorly populated across a set of high quality PDB structures. It became possible in this case, and in other cases not discussed in this work, to identify the more "native-like" loop by introducing a dipeptide-dihedral rotamer frequency-based scoring (RFS) term that penalizes structures with non-native dipeptides confirmations. The details of the RFS will be discussed in a future publication. We applied this penalty term to this loop-hairpin-loop case.

Application of the penalty term ranks the 1.02 Å RMSD prediction lower in energy than the 17.29 Å RMSD prediction (Table X). Aside from 2ZBX, five hairpin cases remain where the predictions remain at around 2 Å or worse. These cases are highlighted in red in Table 4.10. For

these cases, we explored the use of the RFS throughout the entire loop prediction, rather than just to rescore the final loop candidates. The results for these five cases when using the RFS are shown in Table 4.11.

The RFS appears successful at correcting the energy error and leading to a lower RMSD in three of the five cases. PDB 2OKX remains a difficult case. Although this case appears to exhibit an energy error before penalizing unlikely structures with the RFS, now a sampling error remains where we appear unable to produce the native conformation. PDB 3EJA appears to remain an energy error and this case warrants further discussion.

PDB 3EJA contains a 7-residue hairpin within a 15-residue loop that satisfies the various criteria specified in the methods section. In particular the global quality criteria of having suitably highly resolution and superior R-factors was satisfied as well as the local criteria for B-factors and real-space R-factors. Inspection of the predicted loop reveals that we are able to form a reasonable hairpin (Figure 4.10a), and further that during hierarchical loop prediction we succeed in producing a near native loop with an RMSD of 0.94 Å and a ΔE of -1.16 kcal/mol, relative to the native (Figure 4.10b). This would seem to suggest the sampling is not an issue here. The fact that the lowest energy loop predicted (Figure 4.10a) was found nearly 30 kcal/mol lower in energy than the native was surprising. Inspection of the individual residues comprising the loop revealed an unusual close contact between the oxygen on the amide side chain of Q108 and an aromatic carbon on Y191. The distance between these polar and non-polar atoms was a surprising 3.0 Å. Loop minimization perturbs the hairpin such that this distance is increased to 3.5 Å where Y191, like all surrounding residues, is held fixed (Figure 4.10c). The suspicion was that these residues might have been improperly built in the crystal structure and indeed inspection of the electron density showed Y191 to be confidently placed while Q108 was

modeled into sparse density (Figure 4.10d). We see no alternative positions to place Q108, however it is beyond the scope of this work to construct the necessary omit maps and attempt model refinement. In describing the structure, the crystallographers do describe a possible role for Y191 but no mention is made of Q108 and so perhaps this residue simply does not hold a stable conformation. Difficulty in modeling an occasional residue in a high resolution crystal structure is certainly not uncommon. We attempted to exclude loops that were affected by problems such as these in using a real-space R-factor cutoff of 2.0. However, this residue has a real-space R-factor of 0.185. In future studies, it appears a more stringent cutoff is required.

Figure 4.9 Loop-hairpin-loop prediction for PDB 2ZBX. The native loop is shown in gray while the predicted loop is shown in green.

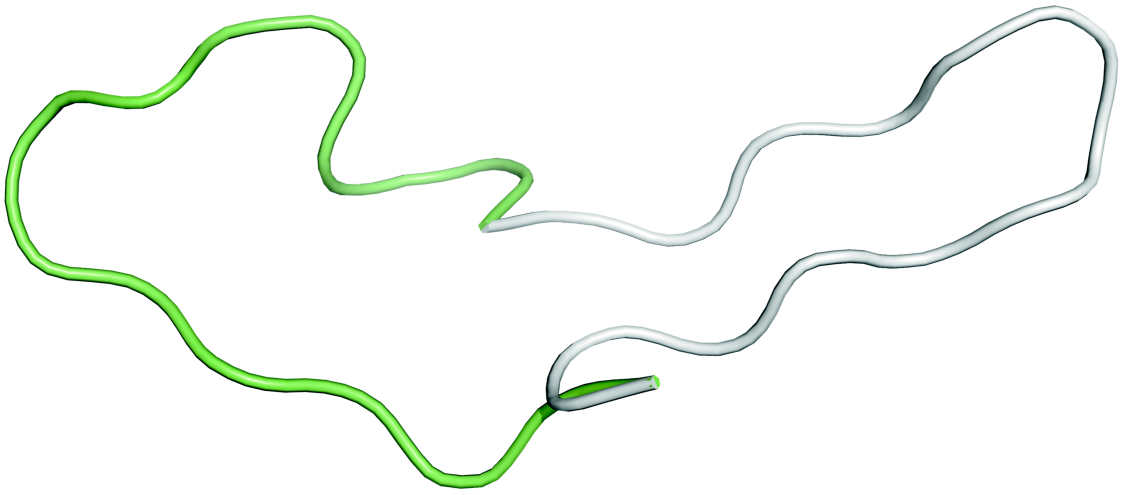


Figure 4.10 Loop-hairpin-loop predictions in PDB 3EJA. In all panels, the native loop is shown in green. A. Native hairpin versus the lowest energy prediction using the RFS. B. Native hairpin versus an intermediately ranked loop. This loop has a 0.94 Å RMSD and a ΔE of -1.16 kcal/mol. C. Native hairpin versus minimization of the native hairpin. After minimization, the distance between Q108 and Y191 increases from 3.0 Å to 3.5 Å. D. 2Fo-Fc map contoured at 2σ around residues Q108 and Y191. Observe that while Y191 is confidently built, Q108 has very poor density.

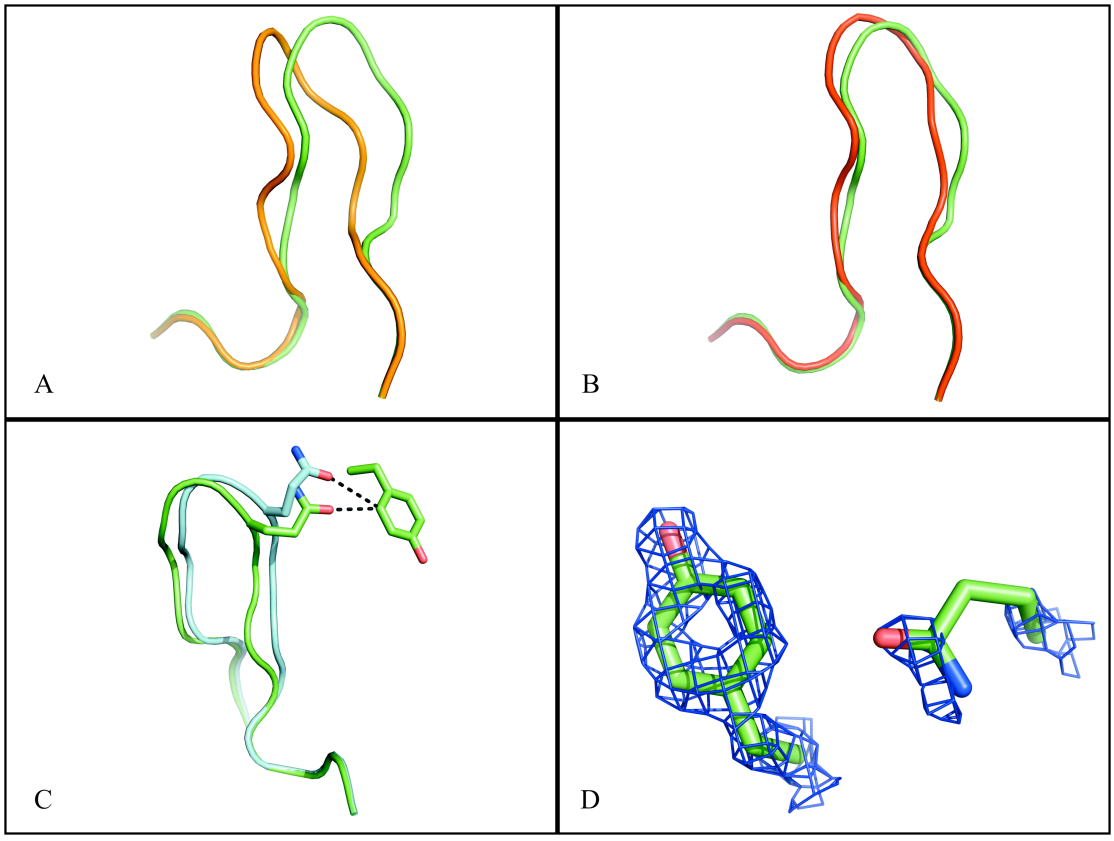


Table 4.8 Results of loop-hairpin-loop predictions using the dipeptide dihedral library. The ΔE value compares the energy of the lowest energy loop against the crystal structure loop coordinates, minimized using our energy function. The RMSD reported is of the lowest energy loop prediction. * - Of the eight-residue hairpins, one of the cases, the loop-hairpin-loop as part of PDB 2ZBX, initially reported the best structure as that with a 17.29 Å RMSD. The results for this prediction were rescored, using the RFS, leading to a 1.02 Å prediction being considered the lowest in energy and was used in the statistics reported in this table. This rescoring is discussed in detail in the text.

Hairpin Length	Number of Cases	Dipeptide Dihedral Library			
		RMSD (Å)		ΔE (kcal/mol)	
		Median	Mean	Median	Mean
6	11	0.41	1.07	-5.61	-5.05
7	2	1.13	1.13	-21.38	-21.38
8*	16	0.64	0.90	-6.47	-6.77
9	7	0.51	0.89	-5.00	-5.74
10	2	0.42	0.42	-7.32	-7.32
11	1	0.53	0.53	-10.55	-10.55
12	1	0.30	0.30	-3.06	-3.06
13	1	0.44	0.44	-0.04	-0.04

Table 4.9 Results of all loop-hairpin-loop predictions. For PDB 2SLI, two hairpins satisfying the criteria described in Materials and Methods were found. Those predictions occurred for the chain A residues 177 - 190 and 236 – 249.

PDB	Loop Length	Hairpin Length	RMSD (Å)	ΔE (kcal/mol)
1C7N	13	8	0.69	-3.34
1F0L	11	6	0.41	-1.71
1GWI	15	8	0.64	-9.05
1GYH	14	9	0.38	-3.18
1LLF	11	6	1.69	-5.61
1NVM	15	9	0.51	-9.55
1O5K	11	6	0.17	-10.79
1TC5	15	8	1.08	1.69
1U60	14	8	0.47	-12.09
1U8V	13	9	0.33	-1.05
2BS2	15	9	0.6	0.03
2C0D	11	7	0.29	-10.5
2CIU	15	10	0.29	-7.11
2IJ2	16	9	0.61	-7.83
2O36	12	9	3.61	-5
2OKX	16	8	2.88	-6.87
2PB2	13	9	0.21	-13.6
2R2N	8	6	0.24	-6.21

2RFG	11	6	0.63	-5.61
2SLI (A: 177 – 190)	14	6	0.26	-0.93
2SLI (A: 236 – 249)	14	8	0.47	-8.88
2WIY	16	8	0.63	-2.36
2WM5	15	8	1.14	-18.43
2YR5	13	6	0.63	-10.95
2YWN	17	13	0.44	-0.04
2ZBX	15	8	1.02	4.70
2ZWA	16	11	0.53	-10.55
2ZYO	8	6	0.36	-1.32
3A9S	12	6	0.18	-4.65
3BF7	11	6	0.98	-9.1
3BJE	12	8	0.34	-3.33
3CSS	17	12	0.30	-3.06
3CU2	11	8	0.38	-3.71
3EGW	12	8	0.49	-10.12
3EI9	15	8	2.12	-2.04
3EJA	15	7	1.97	-32.25
3F8T	14	10	0.54	-7.52
3FAU	13	6	6.21	1.33
3GW9	15	8	0.51	-6.06
3HVW	16	8	0.47	-11.81
3LID	10	8	1.02	-16.62

Table 4.10 Energy of the 2ZBX loop-hairpin-loop predictions after application of the frequency-based penalty term.

RMSD (Å)	Freq.-based Score (kcal/mol)	Total Energy (kcal/mol)	ΔE (kcal/mol)
0.0 (native)	9.89	−15697.1	0.0
1.02	25.65	-15692.4	4.7
17.29	4387.82	−9927.01	5770.09

Table 4.11 Re-prediction of hairpin cases with initial RMSDs of around 2 Å or worse. Re-predictions were performed by using the RFS throughout the prediction, rather than just to rescore the final putative loops.

PDB	**Standard Energy Model**		**Standard Energy Model + RFS**	
	RMSD (Å)	**ΔE (kcal/mol)**	**RMSD (Å)**	**ΔE (kcal/mol)**
2O36	3.61	-5.00	0.93	-10.71
2OKX	2.88	-6.87	3.62	6.42
3EI9	2.12	-2.04	0.36	0.24
3EJA	1.97	-32.25	1.86	-27.2
3FAU	6.21	1.33	0.51	-11.6

4.4.8 Predictions performed in an inexact environment

Throughout all loop predictions, we have relied on the crystal structure to provide the surrounding environment of the loop. This too, like the precise knowledge of helical bounds, may not be accurately known in a homology modeling experiment. To explore the effectiveness of our sampling and energy model in a more realistic setting, we minimized the surrounding environment in the presence of a predicted, but poor, 3 Å RMSD loop. This produced a non-native but locally minimized surrounding side chain environment. However, the backbone environment is still that of the native. From here, we deleted the target loop and performed loop prediction with simultaneous refinement of all surrounding residues. This approach was for both loop-helix-loops and hairpins. We repredicted in an inexact surrounding environment one loop for each secondary-structure length. The loops selected had a sub-1 Å RMSD when predicted in the native environment. For loop-helix-loops, this selection was based on the results from predictions using the exact helical bounds. As would be expected, prediction of the loop as well as surrounding side chains increases the sampling required and computational cost of these predictions. In particular, we found it necessary to introduce additional rounds of side-chain randomization (Table 4.12). Hence, we used only the exact helical bounds to avoid the added complication and expense of sampling surrounding side chains with all the combinations of alternative helical bounds. We also explored the use of the rotamer frequency score (RFS), mentioned previously when describing the improvement in hairpin case 2ZBX (Figure 4.9 and Table 4.10) and others (Table 4.11). Here, we used the RFS throughout the loop prediction, penalizing all intermediate loops as necessary so that only structures with the lowest penalty are likely to advance onto subsequent refinement. The results of these predictions for LHLs are shown in Table 4.12.

In all cases, we were able to recover the loop with sub-1 Å RMSD when utilizing additional rounds of side-chain randomization and the RFS. The use of additional rounds of side-chain randomization finds in all cases a lower energy structure. In 2EX0 the effect is most pronounced where a 2.28 Å prediction is improved to 0.75 Å. Still in the cases 1BKR, 1L5W, and 1WOV, additional rounds of side-chain randomization is further improved with the addition of the RFS, which brings, in the most striking example, a 2.77 Å prediction down to 0.61 Å.

Similar results were seen for hairpins as is shown in Table 4.13. As before, the use of additional rounds of side-chain randomization improves results. Most notably, this additional side chain sampling takes the perturbed native prediction for 2CIU from 6.18 Å to 0.41 Å.

PDB 2C0D evidently posed a significant challenge. The lowest energy structure reported is substantially lower in energy than the native and other similar calculations on 2C0D (Table XIII). This suggests a problem separate from sampling. Visual inspection of the predicted structure relative to the native illustrates the source of this energy error being due to incorrect protonation state assignment.

This situation is illustrated in Figure 4.11. Shown is the close contact between D136 and Y63. Both residues are part of chain B but Y63 is interacting from a crystallographically related monomer. The distance from the carboxylic oxygen in D63 to the C_β is only 3.2 Å while the distance from that same carboxylic oxygen to that residue's backbone carbonyl is 3.35 Å. Were D63 to be assigned as charged, as it originally was using our previously published algorithm(110), substantial repulsion between D136 and Y63 is expected. D136 lies at the tip of the hairpin and so a large deviation of this residue can lead to a significant RMSD for much of the hairpin. Once D136 is assigned as protonated, the successful prediction results. Here, a 0.56 Å loop is produced with a ΔE of -19.02 kcal/mol. The effect of protonation of this residue on all

three perturbed native predictions performed for PDB 2C0D is shown in Table 4.14.

Remarkably, the prediction of this hairpin when the surrounding environment is native is possible with D136 left as deprotonated (Table 4.13). As shown in Figure 4.11b, incorrect protonation state assignment of D136 leads to residue Y63 being perturbed from its native conformation. Evidently, leaving Y63, and all surrounding environment residues constrained to their native position, removes the heavy dependence on correct protonation state assignment of D136. The fact that the removal of this constraint leaves our predictions sensitive to additional factors is not surprising. Additional perturbed native experiments such as these will be run in the future to expose more weaknesses in our algorithm, however for the cases presented in this work, the difficulties appear isolated to this case and are tractable.

Figure 4.11 PDB 2C0D. In all panels, the native loop is shown in green for comparison. A: The native loop with all atoms shown for D136 and surrounding side chains Y63 and T64. The suspicious close-contacts that motivated protonation of D136 are shown dotted in this panel. B: The coordinates of the same atoms in the RFS prediction with D136 deprotonated. C: The coordinates of the RFS prediction with D136 protonated. Notice the similarity to the native loop in panel A.

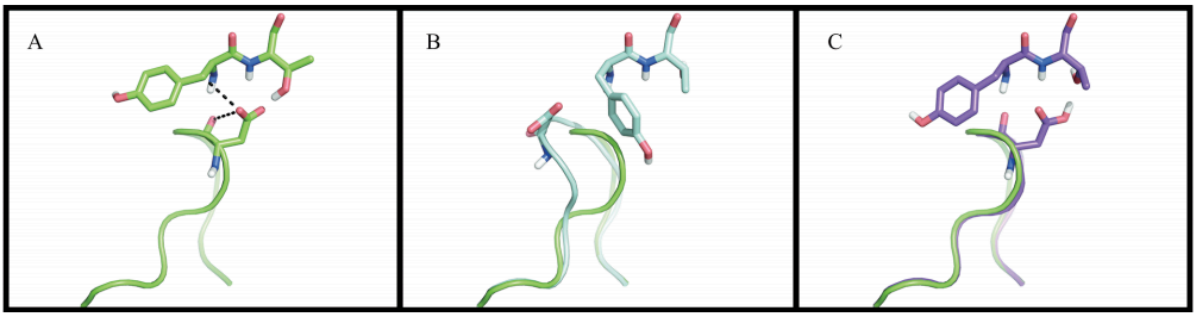


Table 4.12 Results from LHL prediction in an inexact environment. The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized.

Helix Length	PDB	Native Environment		Perturbed Native		Perturbed Native with extra side-chain randomization		Perturbed Native with extra side-chain randomization and RFS	
		RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)
4	1BKR	0.55	1.05	1.67	24.54	2.77	2.42	0.61	-2.11
5	1L5W	0.4	-3.74	0.78	-1.39	0.98	-8.97	0.54	-15.03
6	1WOV	0.67	-2.88	1.29	5.25	1.32	-12.85	0.66	-22.55
7	3HL0	0.39	2.52	0.62	-6.97	0.6	-16.28	0.68	-17.84
8	2EX0	0.33	-4.40	2.28	23.84	0.54	10.49	0.76	7.77

Table 4.13 Results from hairpin prediction in an inexact environment. The RMSD is relative to the native structure. The ΔE shown is relative to the energy of the native where the loop and surrounding side chains are minimized. The hairpin of length 7, 2C0D is shown before protonation of D136 in chain B. After protonation of this residue, the energy errors shown here are eliminated. Energy errors occur when predicted loops are reported substantially lower in energy than the native but have poor RMSD. This is discussed in greater detail in the text.

Hair-pin Length	PDB	Native Environment		Perturbed Native		Perturbed Native + addl. side-chain randomization		Perturbed Native + addl. side-chain randomization + RFS	
		RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)	RMSD (Å)	ΔE (kcal/ mol)
6	1F0L	0.41	-1.71	0.72	12.68	0.74	-10.01	0.73	-14.16
7*	2C0D	0.29	0.89	0.89	-14.19	2.34	-27.39	1.71	-1.54
8	2SLI	0.48	-6.85	0.54	-1.64	3.22	-8.67	0.49	-12.72
9	1GYH	0.38	-3.18	0.73	0.45	0.82	0.18	0.9	1.54
10	2CIU	0.29	-7.11	6.18	29.67	0.41	-22.61	0.57	-10.21
11	2ZWA	0.53	-10.55	0.91	-10.16	0.46	-6.17	0.77	7.68
12	3CSS	0.30	-3.06	0.57	2.05	0.4	-4.86	0.37	-3.73

Table 4.14 The effect of protonation of D136 on the hairpin prediction in PDB 2C0D.

D136 Protonation State	**Perturbed Native**		**Perturbed Native + addl. side-chain randomization**		**Perturbed Native + addl. side-chain randomization + RFS**	
	RMSD (Å)	**ΔE (kcal/mol)**	**RMSD (Å)**	**ΔE (kcal/mol)**	**RMSD (Å)**	**ΔE (kcal/mol)**
Deprotonated	0.89	-14.19	2.34	-27.39	1.71	-1.54
Protonated	1.34	19.14	0.71	-22.56	0.56	-19.02

4.4.9 Interpretation of the relative energies

Throughout this work, we have reported results comparing the geometry of our predicted structure to the native coordinates via the RMSD, and comparing the energy of our predicted structure to the minimized native via the ΔE. As mentioned in the methods section, $\Delta E = E_{prediction} - E_{native}$. In any successful energy model, the minimized native structure should be reported as being lowest in energy and yet we report negative ΔE values across various predictions. It is worth speculating on the source of this. We believe there are two general possibilities:

1. There are problems in the backbone of the crystal structure that cannot be rectified with our gradient-based minimization as our energy model places the backbone in a local minimum. This seems perfectly plausible in crystal structures, even for the high quality structures explored in this work, as hydrogen atoms positions are not experimentally known, preventing, at the least, the use of an all-atom energy model for refinement. Indeed, Bell *et al.,* report a successful reduction in non-bonded clashes in crystal structures, introduced after consideration of explicit hydrogen atoms, through the use of an all-atom refinement procedure without any loss in adherence to the diffraction data(119). Thus, what we may be observing instead is a slightly physically superior structure obtained during the extensive sampling performed during our *ab initio* loop prediction.
2. That negative ΔE values observed in predictions with remarkably low sub-Ångström backbone RMSD may instead be due to improper side-chain contacts being formed. For example, Table 4.12 includes a 0.33 Å prediction of an LHL in PDB 2EX0 with a ΔE of -4.40 kcal/mol. It may well be that these improper contacts are due to a flaw in our

energy model, and although this is possible, our ability here to select the lowest energy structure and achieve sub-Ångström RMSDs appears unaffected. As such, in this paper we do not investigate in greater detail the source of these errors.

We also observe systematic differences in the ΔE across methods and secondary structure. For example, Table 4.1 reports the RMSD and ΔE of LHL predictions performed using just our normal dipeptide dihedral library versus the helical dihedral library presented in this work. In this table, the mean ΔE for all helix lengths predicted is lower with the helical dihedral library than without. This suggests that without the helical dihedral library, there are sampling errors which are removed by seeding the helix.

For the hairpin predictions, Table 4.13 and Table 4.9 show that the vast majority of predictions conclude with a structure with a negative ΔE. Referring to the first of our two speculations on the source of these negative ΔE values, it may be that the extensive sampling performed in loop prediction is producing superior backbone hydrogen bonds that are not accessible through minimization of the crystal structure.

4.5 Conclusions

We have developed a robust algorithm to exploit secondary structure prediction of small helical segments in loops to yield routinely accurate loop-helix-loops predictions to atomic accuracy. Furthermore, we have demonstrated that our previous dipeptide-dihedral library and all-atom energy model can successfully predict loops containing hairpins. By running parallel loop predictions with a systematically generated set of putative helical bounds from two secondary structure prediction algorithms (SSPro4 and PSIPRED) as well as the normal loop prediction protocol, we have demonstrated that the native loop-helix-loop can be reliably sampled and accurately scored.

This application of a separate, helical dihedral library to a subset of loop residues is at the crux of our method. It affords us increased likelihood of the formation of the coupled hydrogen bonds that define secondary structure by performing loop buildup with the coupled dihedral angles already in place, but it has also introduced a sort of lever effect, where small changes at the base of the helix lead to significant displacement of the terminal end of the loop. For smaller helices, this is obviously less of a problem but for larger helical bounds, such as the LHLs predicted in PDBs 1OAO and 2YR5 where the helical bounds were supplied by PSIPRED, it became impossible for loop buildup to be performed – all possible helix conformations produced loop halves that were considered impossible to close.

Rather than seek to expand the size of our helical dihedral library to include more rotamers, we found it more effective to attempt loop-helix-loop prediction with shorter helical bounds, one that would be less likely to demonstrate a lever effect. This led to the use of our truncated helix sampling method. We leave it up to subsequent rounds of further minimization and sampling to form the remainder of the helix, and indeed this appears to be effective. Nonetheless, for very large helices, our limited dihedral library may fail to contain a sufficient number of rotamers to avoid a sampling error and the truncation method may leave too large of a sub-loop to correctly sample and form the remaining coupled dihedrals that are necessary to complete the helix. In practice though, this is not a very large concern for us. Such large helices are likely the well-conserved regions between homologous proteins. Knowledge of these helical bounds would likely be found with sequence-based secondary structure prediction methods, but crucially, the conformation of these large loop-helix-loops lies squarely within the purview of our previous rigid helix placement algorithm(99).

Hairpins, somewhat surprisingly, appeared as a simpler type of secondary structure to

predict. The small non-locality of the hydrogen bonds deterred us from wanting to introduce a separate hairpin dihedral library as such a library would seem to produce a bias in the non-hydrogen bond turn-region of the hairpin between the two β-strands. Rather, we attempted loop-hairpin-loop prediction using only our previous dipeptide-dihedral library(104). Low RMSD loops were successfully predicted to atomic accuracy with no significant change to our past algorithms, other than permitting a flexible *ofac* to be tried throughout all rounds of hierarchical loop prediction. For both hairpins and loop-helix-loops, it would be desirable in the future to further establish this methodology by running blind tests where the structure of a given loop is available but unknown to the researcher. However, we do not anticipate the results of such experiments to diverge from what we present here as our method is automated, using only the energy and not user input, to determine the final loop conformation.

Predictions performed in a non-native surrounding environment were successful, albeit requiring additional sampling and the use of our rotamer frequency score to accurately predict the loop. An apparent caveat is that the additional degree of freedom now present in the surrounding environment can magnify energy errors. As shown in the hairpin in PDB 2C0D, incorrect protonation state assignment of an aspartic acid is compensated for through the coupled movement of a surrounding environment residue. Although only this case had such a problem, clearly more experiments need to performed across a large set of loops, with and without secondary structure, to expose weaknesses in our algorithm and correct them. These experiments are already underway and will be discussed in a future publication.

Chapter 5

Docking into the Kappa opioid receptor

5.1 Note

This project emerged when Prof. James Leighton came to us and proposed that we help with a drug discovery project on the kappa opioid receptor. Upon discovering that the existing scoring functions within the docking program GLIDE failed to pick out actives from decoys we decided to try the newest scoring function that works with glide, named WScore. WScore is still in development and to date continues to be devised and tested. The nature of WScore development is also slow, because every change made must be tested for generality against all receptor-ligand complexes in the training set. This included new terms that came out of working with the Kappa opioid receptor. At this point, our results paint a complete picture, thus I am including it in my thesis. However, we have a few other ideas and calculations to implement and run, respectively, making this chapter an overview of what has been accomplished thus far. The full set of results will be submitted for publication soon, with additional details and analysis.

5.2 Introduction

5.2.1 The Kappa opioid receptor

Opioids are the most widely prescribed and abused pharmaceuticals, due to their powerful painkilling, antidepressant, and addictive properties (120, 121). Opioids also have severe side effects ranging from constipation to dysphoria, which has motivated discovery of safer and nonaddictive medications (121). Manmade opioids bind promiscuously to the mu, delta and kappa opioid receptors (MOR, DOR, and KOR, respectively), and it is known that MOR is involved in the addiction pathway (122-126).

Thus, it is hoped that high-affinity ligands that selectively bind to KOR or DOR might be the key to finding nonaddictive opioids. KOR agonists have demonstrated desired pharmaceutical effects, while avoiding activation of reward pathways, although to date they do not avoid adverse effects like dysphoria. Fortunately, dysphoria appears to be triggered by arrestin recruitment to activate the receptor, implying that selective KOR ligands could be devised that avoid this pathway (127-129). Like all structure based drug design efforts, however, doing so requires a detailed molecular picture of how opioids bind to their receptors.

Such a picture is now possible as the kappa opioid receptor was crystallized in 2012 bound to the antagonist JDTic (3R)-1,2,3,4-tetrahydro-7-hydroxy-N-[(1S)-1-[[3R,4R)-4-(3-hydroxyphenyl)-3,4-dimethyl-1-piperidinyl]methyl]-2-methylpropyl]-3-isoquininolinecarboxamide) molecule (32). We also have access to the three other main opioid receptors—MOR, DOR, and the nociceptitin/orphanin FQ peptide receptor—which, like KOR, belong to the class A gamma subfamily of GPCRs (130-132). True to the common GPCR architecture, they contain seven transmembrane helices connected by alternating intra and extracellular loops. The KOR, DOR, and MOR subtypes are highly homologous, exhibiting around 70% sequence identity in the transmembrane region, which houses their binding sites (133, 134). From these crystal structures, one can hope to develop a molecular understanding of how the opioid receptors discriminate between active and nonactive ligands, why many of the same molecules, including morphine, bind to them, and with what relative binding affinity.

With this in mind, we used the new KOR crystal structure (and MOR and DOR for comparison) to construct a theory of how the receptor binds morphinan antagonists and,

with some rearrangement of the active site, agonists. Our theory is based on the presence of structurally fixed waters within the binding site and the increase in binding free energy that comes from displacing them in a favorable way. This limits what types of ligands can bind and elucidates the binding mechanism of KOR. This should prove useful to researchers searching for novel KOR binding ligands. We integrated the key elements of the theory as new terms to a new scoring function, which allows us to pick out known active antagonists from a set of chemically similar decoys and provides experimental validation of the method.

5.2.2 Overview of the WScore scoring function

The Wscore scoring function is a major revision of the Glide XP scoring function (9-13) in which the localized water structure in the active site, defined by a WaterMap molecular dynamics simulation, is integrated into the Glide XP model. Within the rigid receptor approximation, the position of the WaterMap water sites is rigorously defined on the Glide XP docking grid, as is the (approximate) displacement free energy of these waters. The interaction of the ligand with the water structure can then be specified much more thoroughly than in a conventional empirical scoring function, in which only the interactions of the ligand with the protein are considered explicitly. An initial summary of the Wscore approach, along with preliminary results for several targets using the DUD data set, has been presented in *Repasky et al* (9).

Glide XP contains a number of terms which delineate specific molecular recognition motifs, including recessed salt bridges, displacement of hydrophobically enclosed water, and the formation of correlated hydrogen bonds in hydrophobically enclosed regions. These terms are highly effective in identifying the principal drivers of potency in known

active compounds. However, as discussed in refs 9-13, Glide XP as originally formulated lacks terms to identify strain energy, desolvation, and trapping of water in a hydrophobic pocket by the ligand. Wscore remedies the first of these deficiencies using geometrical criteria for assigning high energy ligand rotamer states, and the second by using interactions of the WaterMap waters with the ligand to assign desolvation penalties and penalties for water molecules which are destabilized by hydrophobic contacts of the ligand.

An important point is that the Wscore desolvation and ligand strain terms are not receptor specific; they are global terms, dependent upon the geometry of the protein-ligand complex and associated WaterMap waters, which can be applied to an arbitrary ligand and receptor. All penalty terms that have been implemented have been tested on a large data set of ~600 protein-ligand complexes from the PDB, and avoid applying penalties to these complexes which would inappropriately reduce the score of known active compounds. Furthermore, the various terms in the model are required to be physically motivated, as opposed to arising from fitting a large number of descriptors to experimental structure-activity relationships.

In the present discussion, we focus on two key terms in Wscore which play a critical role in discriminating active compounds from decoys in the KOR docking experiments that we have carried out. Firstly, it can be noted that the great majority of known KOR tight binders, with the exception of salvinorin, form a salt bridge with Asp 138, using a positively charged nitrogen for this purpose. Formation of this salt bridge is necessary in these cases to avoid desolvation of Asp 138 by the ligand. We describe the WScore desolvation term which enforces this condition below. Secondly, in our analysis of the WaterMap derived water structure in the KOR active site, we have uncovered a highly unusual water structure

which we hypothesize plays a novel role in KOR molecular recognition. No analogue of this water structure is found in any of the other complexes in our PDB derived test suite, although similar structures may well be present in other GPCR active sites. We implement a simple but effective model to describe the effects of this water structure on KOR ligand binding, and it provides a readily understandable mechanism for discriminating active compounds from decoys in many cases.

5.3 WScore discussion

5.3.1 Desolvation of charged residues

A substantial fraction of charged residues in proteins place their side chains in highly solvent exposed locations, such that their solvation free energy is comparable to what is achieved in bulk solution. An alternative conformation with acceptable free energy is the formation of a salt bridge with a complementary protein residue, typically one with a significant degree of solvent exposure. Salt bridges which are recessed (i.e. have restricted solvent exposure) are also observed, typically when there are no good alternatives.

When a ligand group is proximate to a charged residue in bulk solution, or in a salt bridge, there is generally not a highly deleterious effect on binding free energy, even if the group is hydrophobic. In bulk solution, water molecules can reorganize around the charged atoms and construct a reasonable alternative first solvation shell. If the group is in a salt bridge, solvation free energy of the charge pair is significantly smaller than that of an isolated charged residue (as the electric field is essentially now dipolar), so again, perturbations of the first shell can be tolerated as long as they are not too extreme.

However, there are cases in which a charged residue is positioned in the active site in a confined space, without making a salt bridge. WaterMap calculations on such

structures frequently reveal multiple localized waters which are hydrogen bonded to the charged side chain. Such a water structure indicates that satisfactory solvation of the side chain requires a specific water structure, as opposed to the situation in bulk solution where there are a large number of different solvent configurations with similarly favorable interactions with the charged group. Disruption of this water structure by the ligand, without forming a compensating salt bridge to the side chain, is thus much more likely to lead to a substantial loss of free energy. This hypothesis is confirmed by examination of a large range of complexes in the PDB. There are a very small number of complexes in which a single WaterMap water bound to a charged side chain is displaced without compensation, but none at all in which more than one such water is displaced without compensation. Therefore, a major penalty term is assigned when a ligand causes displacement in this situation.

In the KOR active site, Asp 138 is buried in the pocket and does not form a salt bridge with another protein residue. Examination of the WaterMap results reveal that Asp 138 is bound to 7 WaterMap waters (see Fig. 5.1a). Displacement of these waters without forming a salt bridge leads to a structure in which the charge on the Asp is buried beneath the ligand. The penalty term described above enforces this constraint. In the KOR crystal structure, JDTic forms two salt bridges to Asp 138 (see Fig. 5.1b). The salt bridges compensate for the fact that the ligand displaces all but one of the formerly Asp 138 stabilizing waters (see Fig. 5.1c). The importance of the salt bridges was identified by the authors of the crystal structure paper as JDTic contains two protonated amines in the piperidine and isoquinoline moieties that form salt bridges. D138A mutagenesis experiments also demonstrated an almost 100-fold reduction in binding affinity.

Experiments showed that if the isoquinoline nitrogen is replaced by a carbon, oxygen or sulphur atom, the binding affinity is not significantly reduced, alluding to the fact that a single salt bridge is sufficiently stabilizing that binding to the receptor is not hindered.

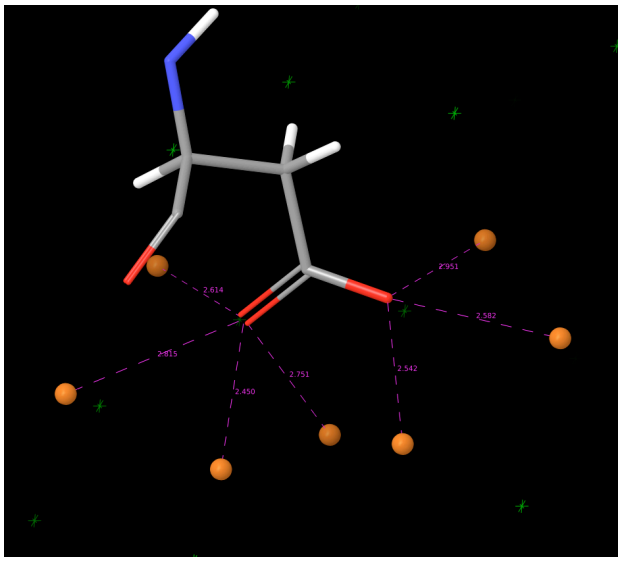
To get an accurate picture of water inside protein active sites, much higher resolution than the 2.9A of KOR is needed. Nonetheless, the authors do see experimental evidence of structured water molecules near both distal hydroxyls on JDTic. The WaterMap structure explains why the vast majority of known active compounds binding to KOR contain a positively charged nitrogen which can form a salt bridge with Asp 138. There are a few known compounds, such as salvinorin A, which bind effectively to KOR but do not contain a positively charged nitrogen. The binding mode of this compound must be quite different from the typical KOR strong binder, many of which have structures (and presumably binding modes) similar to morphine (and, naturally, other morphinans), or the antagonist in the crystal structure (PDBID 4DJH), JDTic. It is quite possible, even likely, that the active site of KOR requires substantial reorganization from the 4DJH structure in order to tightly bind salvinorin A, and hence salvinorin A binding cannot be modeled accurately using rigid receptor docking into 4DJH.

In the results presented below, we shall see that all compounds achieving good WScore rankings form a salt bridge with Asp 138, for the reasons given above. Some of the known active compounds with a positive nitrogen fail to do this, presumably because treatment of induced fit effects (which we do not consider in this paper) is required to achieve the necessary geometry. A number of the active compounds in this category are very large and may require expansion of the binding site. Others have a chemical structure that is very different from the morphine or JDTic analogues in the data set, and may need a

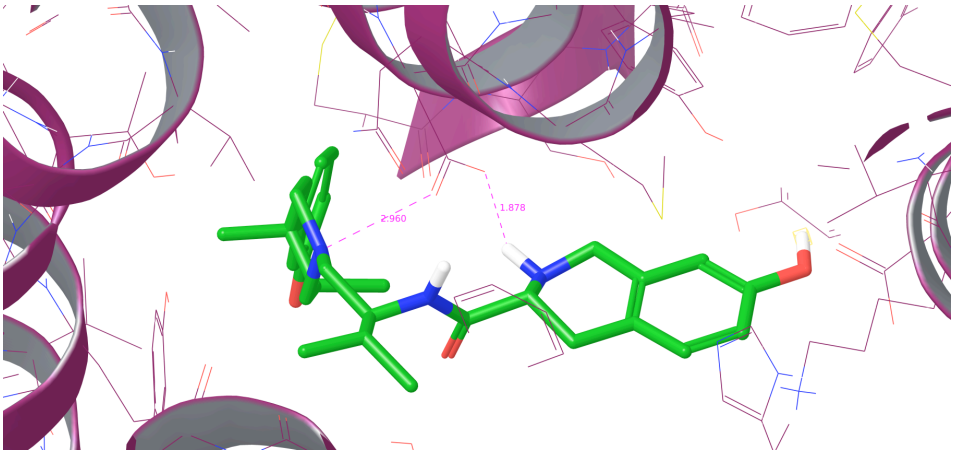
very different set of side chain conformations to bind in the correct conformation. These issues will be considered in detail in future work.

Figure 5.1 Salt bridge to Asp 138. a. 7 waters (in orange) solavting Asp 138. b. salt bridges forming between JDTic and Asp 138. c. 6 water (in orange) displaced by JDTic, leaving only 1 remaining water (in yellow), which would leave Asp 138 desolvated if not for the compensating salt bridges formed with basic amines.

a.

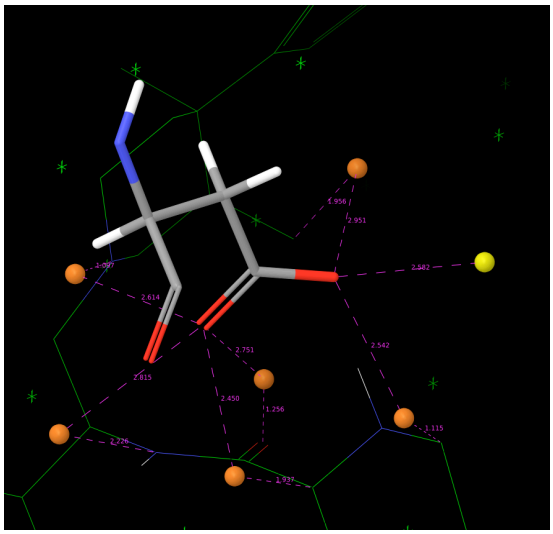


b.



.

c.



5.3.2 Water structure in the KOR receptor: effects on the WScore scoring function

The active site of KOR (as well as the mu (MOR) and delta (DOR) opioid receptors, and likely other GPCRs) contains a large number of water molecules. In our WaterMap simulation, these water molecules are for the most part localized, leading to a network of water molecules forming multiple hydrogen bonds, as in the structure of ice.

A WaterMap water site is distinguished by a number of features; firstly, the approximate prediction for the free energy of displacement, and secondly, the number of hydrogen bonds that are formed to neighboring water molecules and protein polar or charged groups. Ordinarily, when a waters site makes four hydrogen bonds, a significant number of these hydrogen bonds are with protein groups, and the water is difficult to displace in terms of free energy. However, the KOR WaterMap reveals one water molecule (termed water 1), shown in Fig. 5.2a, which makes four hydrogen bonds, 3 with other WaterMap sites and one with Tyr 139, and has an extremely unfavorable (i.e. easy to displace) free energy (~6 kcal/mole). Such a site resembles that of a water molecule in ice

at room temperature, where the low entropy of the site leads to a very unfavorable free energy. Thus, in this region of the KOR active site, there is an unusual degree of structuring of the water molecules despite the lack of immediately proximate protein groups. In other receptors, binding pockets are generally smaller and narrower, so that the possibility of forming an ice-like structure is limited. In the present case, the combination of significant confinement of the water, coupled with the lack of protein groups for water water 1 to hydrogen bond to, leads to the unusual situation described above.

The local structure in this region appears even more ice-like when one considers the fact that water 1 is hydrogen bonded to water 2 (see Fig. 5.2a), which is also an extremely unstable water (~6 kcal/mole). No such pair of waters can be found in any of the ~600 PDB complexes that make up the standard Wscore training set. Water 2 makes only three hydrogen bonds, one to water 1, one to another WaterMap water, and one to Asp 223.

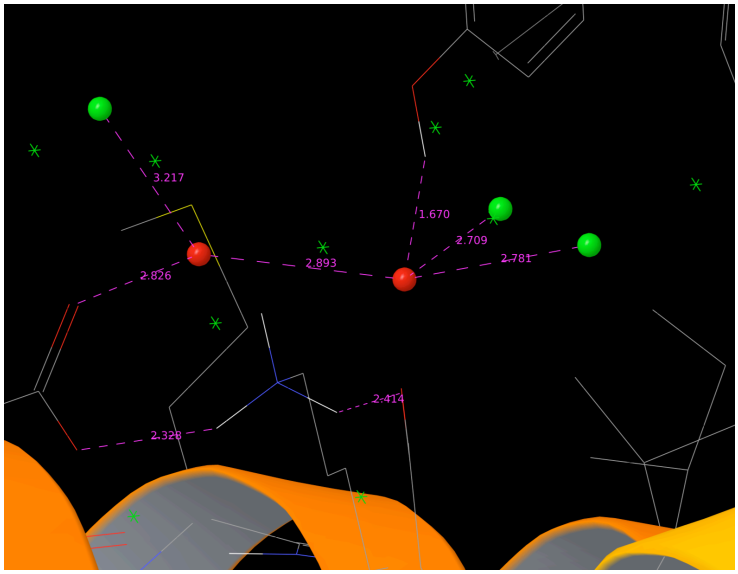
The significance of water 1 is suggested in **(32)** (there is evidence of it being a crystal water in the structure), and becomes manifest when examining the crystal structure of JDTic, or the docked poses for morphine-like actives docked into 4DJH. In the JDTic structure, the hydroxyl bound to the isoquinoline moiety makes a hydrogen bond to water 1, compensating for the 2 displaced waters previously bonded to water 1 (see Fig. 5.2b). If the hydroxyl is mutated to a hydrogen, 3 kcal/mole of binding affinity is lost (ie. a 100 fold reduction in affinity). In morphine and actives with a morphine-like chemotype, two hydrogen bonds are made to water 1, and this appears to lead to additional binding affinity.

Ordinarily, water-mediated hydrogen bonding is not a significant driver of potency. However, when a water molecule has the extraordinary, ice-like structure of water 1, loss of a hydrogen bond to another water will be very costly, because the hydrogen bond energy

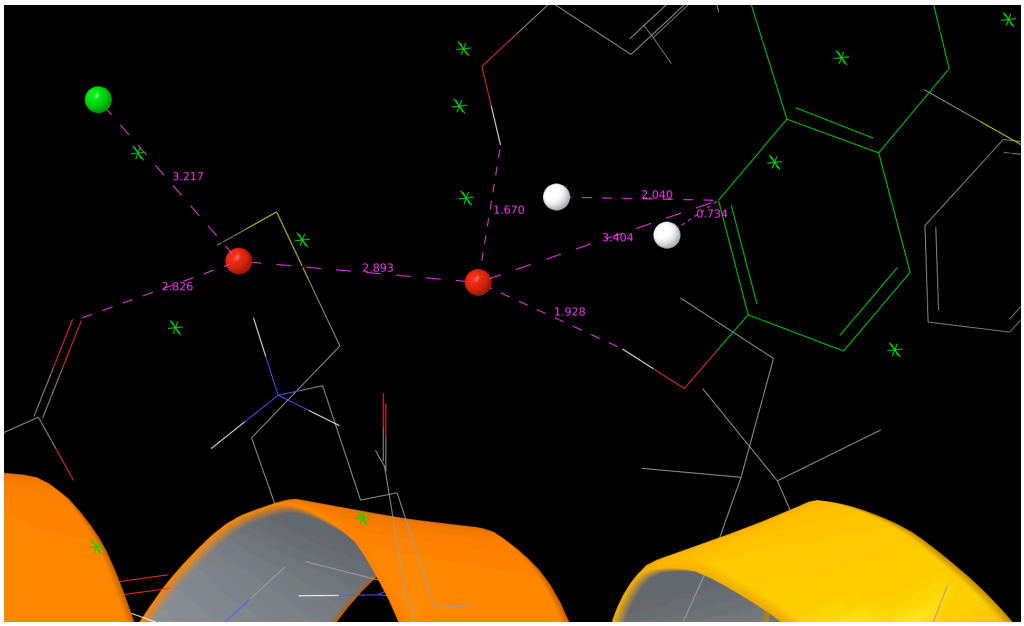
lost cannot be made up by either forming an alternative hydrogen bond in a different geometry, or via an increase in entropy. Hence, we apply a penalty to docked poses in which water 1 loses hydrogen bonds which are not replaced by a sufficiently favorable interaction with the ligand. Based on prior experience with WScore parametrization, and taking into account the unusual features of the present structure, the expectation would be that replacement of a hydrogen bonded water with a hydrogen bond from the ligand would result in a favorable free energy gain (as the displaced water gains entropy by going into bulk solution), whereas replacement of a hydrogen bond with an aromatic C-H bond to the water would result in an unfavorable free energy change (as the aromatic C-H bond is not as strong as a normal hydrogen bond, and the hydrogen bond strength is critical here due to the relative immobility of water 1). As many of the decoy molecules displace multiple waters surrounding water 1 but fail to form any hydrogen bonds with water 1, this term leads to significantly improved discrimination between actives and decoys. It also helps in explaining the very tight binding of morphine, which can make 3 hydrogen bonds to water 1, despite the small size of this molecule.

Figure 5.2 High energy water pair. a. water 1 (rightmost red) is bound to water 2 (leftmost red), Tyr 139, and 2 other water (rightmost greens). Water 2 is bound to water 1, Asp 223 and antoher water (leftmost green). b. In the presence of JDTic, the two waters bound to water 1 (in white) are displaced, but a compensatory hydrogen bond as well as an aromatic CH (both labeled) are made.

a.



b.



5.4 Results using WScore

Having integrated our insights from the water structure within the KOR active site into WScore, we then attempted to verify our model. To begin, we redocked JDTic into KOR (see Fig. 5.3). The docked pose has an RMSD of 0.86A to the position in the crystal structure and a score of -11.0, which corresponds to about 1nm binding. The experimental binding affinity of JDTic to the crystallized KOR with a T4 lysozyme inserted at the site of intracellular loop 3 is 0.6nM. Docking JDTic without the two distal hydroxyl groups yields a score of -7.8, or roughly a 3kcal/mol difference with the native ligand, which corresponds to about a 100-fold reduction in binding affinity and quantitatively matches experiment. The docked poses of the JDTic with and without the hydroxyl groups are the same, meaning that the lower score of the JDTic without hydroxyls is not coming from misdocking it.

Redocking a receptor's cocrystallized ligand is a necessary but not sufficient criterion to be able to pick out cross docked known active ligands from a set of decoys. It is well known from the literature that ensemble docking gives the best crossdocking results—a natural consequence of the fact that even similar looking ligands can require a large shift in the active site. There is only one crystal structure of KOR in the antagonist bound state, and many active antagonists will fit into the crystal structure. Thus, the terms added to the scoring function have to be flexible enough to accommodate ligands that require modest reorganization of the binding pocket; more serious reorganization will require new crystal structures. Induced fit docking methods can also be tried. Using the Cavasotto GPCR dataset for KOR, we docked 27 known active antagonists and a set of decoys that ensure ligand-decoy similarity of six physical properties while also enforcing chemical dissimilarity. This provides a challenging test set for WScore as the decoys are

similar enough that they are rarely unable to find any suitable spot in the binding site. All but four of the 27 actives contain a morphinan core.

The results show excellent early enrichment (see Fig. 5.4a and 5.4b). 26% of the actives appear the top 2.7% of all docked structures (actives and decoys), the top scoring active is only outranked by 5 decoys, and are all morphinans. They overlay the binding mode of JDTic, and stabilizing the high energy water pair with hydrogen and aromatic CH bonds. The binding mode also parallels that seen of the cocrystallized morphinans in MOR and DOR, lending further experimental evidence to the computational docking results. In contrast, the previous SP and XP scoring functions yielded zero actives in the top 2% of docked structures, with 76 and 66 decoys outranking the top scoring active, respectively.

Of high scoring decoys, the vast majority fall into one of two groups: they either appear to bind in the mode we expect and look as if they could be active, or they overlay the piperidine moiety in JDTic. At the present time, we can only speculate as to whether or not the binding mode that overlays the piperidine moiety of JDTic is valid. We are planning to run rigorous free energy perturbation calculations very soon to answer this question. The decoys whose binding modes follow the binding mode that we believe to be correct may indeed bind to the KOR. As emphasized before, the decoy set is based on known binders, and it would not be surprising if a few of them are actually micromolar binders. The only way at this point to determine whether or not they do actually bind to the receptor would be to run binding assay experiments.

The other morphinan active antagonists that do not score well appear to mostly be misdocked due to induced fit effects (as discussed before). Two docking with the correct binding mode, but still do not fit well into the precise shape of the binding pocket. A good

indicator of bad fit in a binding pocket is unfavorable contacts. They are reflected in a lower Van der Waals energy component. As seen in Fig. 5.5, the worse scoring actives have lower trending Van der Waals energies. The remaining actives in this dataset are not morphinans, one docks and appears to interact favorably with the high energy water pair, while the other three lie in a different part of the active site. We cannot be confident in this binding modes as there is no experimental validation for these types of molecules.

Overall, this study emphasizes the importance of the crystal water structure within an active site. Generally speaking, displacing unstable waters which are hydrophobically enclosed or solvate free charges increases the free energy of a system, as long as the ligand replaces the water with hydrophobic groups or appropriate ligand-receptor hydrogen bonds, respectively. Without these explicit waters present in the docking process, it is much harder to estimate such energy gains from water displacement. Equally important, the displacement of unstable waters without a ligand forming more favorable interactions is very bad. WScore penalizes this situation, and it is these penalties that lie at the heart of the success of the scoring function.

The MOR crystal and watermap water structure within the binding site show a single high energy water where the pair is located in KOR's active site (see Fig. 5.6). It is even more isolated that the pair, being only hydrogen bound to a tyrosine (that parallels Tyr 139 discussed before), and 1 other water, which is displaced by the cocrystalized morphinan ligand. This would leave the high energy water in a very unfavorable energetic state except that it can now hydrogen bond to an oxygen and a hydroxyl located on the ligand. This similarity in morphinan binding across the opioid receptors serves as an explanation for why so many known ligands bind to all three.

Figure 5.3 The docked lowest energy pose of JDTic (in blue) overlays the crystal structure of JDTic bound to KOR.

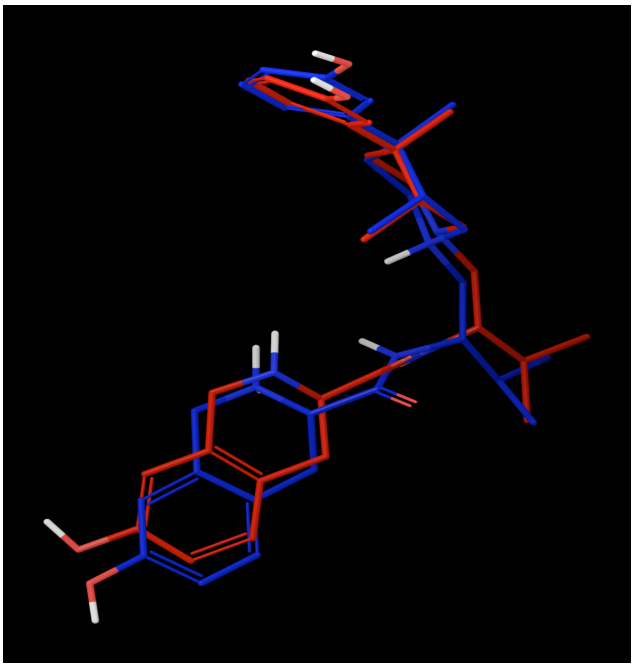
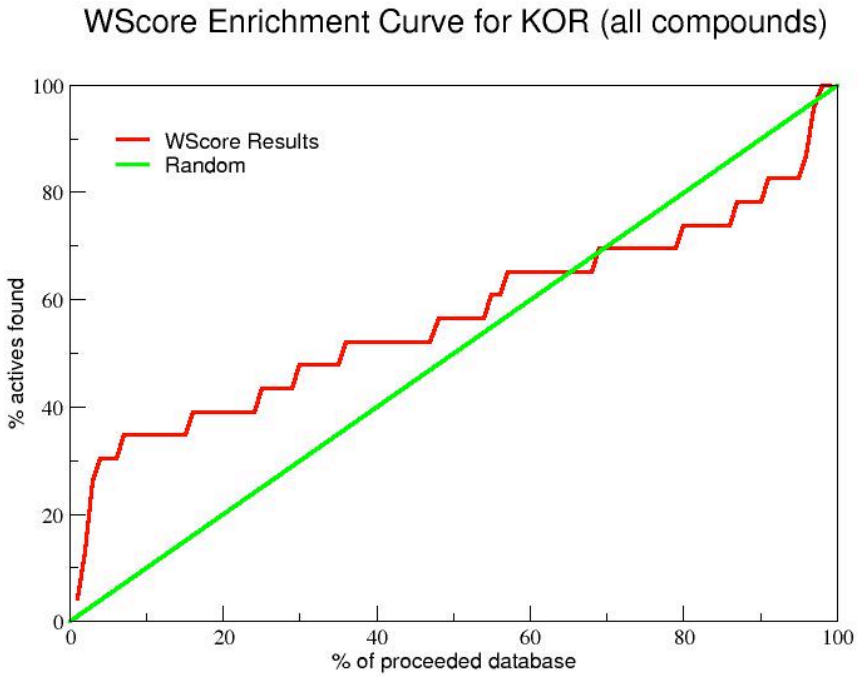


Figure 5.4 Enrichment curves. a. The WScore enrichment curve for KOR for all actives and decoys. b. The WScore enrichment curve for KOR for the top 2.7% of actives and decoys.

a.



b.

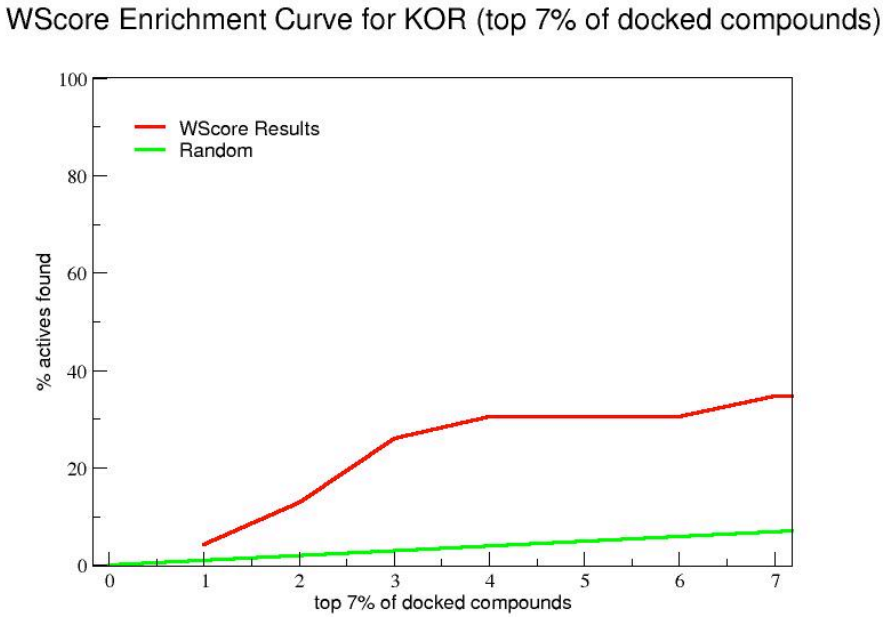


Figure 5.5 Actives that do not fit in the receptor tend to have smaller Van der Waals Energy, as seen by this curve. The black curve includes all of the actives, while the pink curve excludes two unusual actives that contain chlorines.

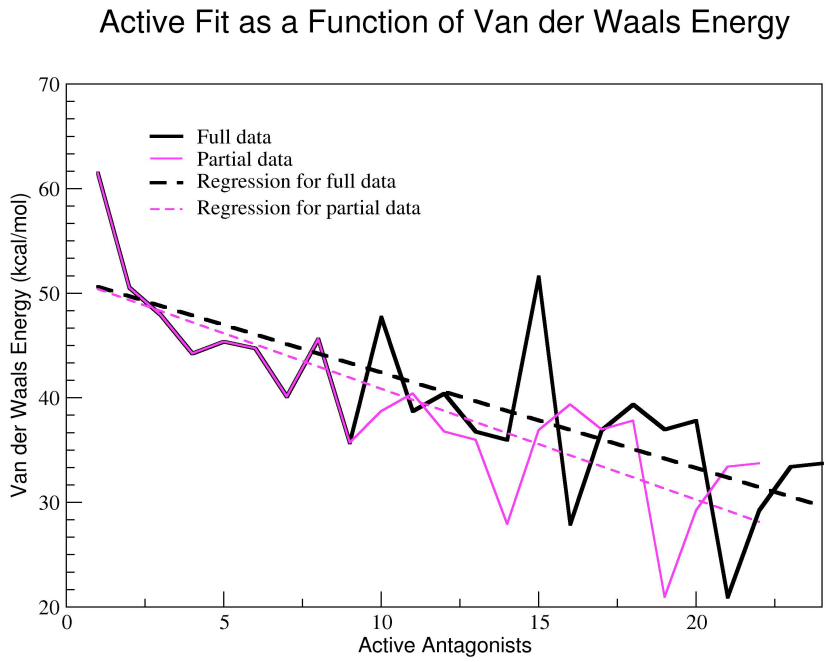
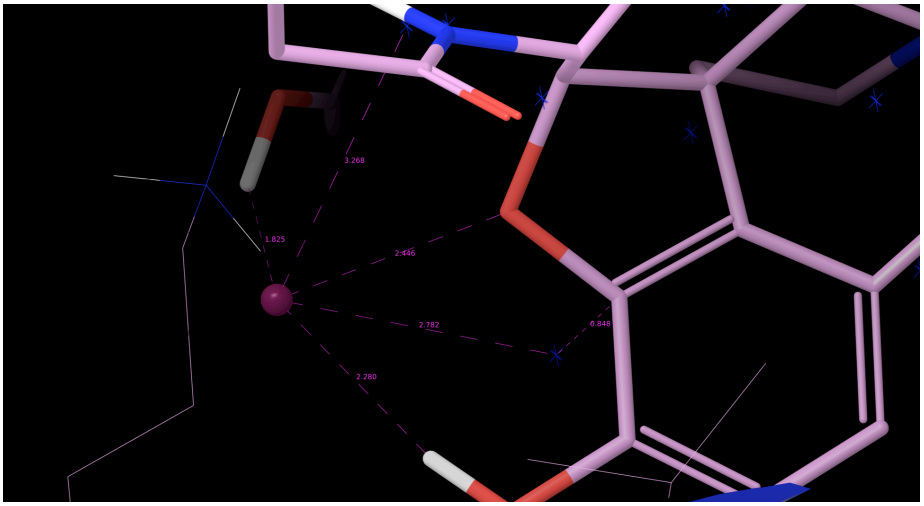


Figure 5.6 The cocrystalized ligand in MOR and the hypothesized importance of one high energy water molecule (in purple) to its binding mode.



5.5 Conclusions

The formation of a salt bridge being key to ligand binding for the KOR appears to follow other GPCRs, but like most receptors, there is a second piece of the binding regulation puzzle: a high energy water pair that must be kept stabilized or else a ligand cannot favorable bind. This picture is supported by experiment but for the first time understood with a detailed molecular explanation. Furthermore, we now have a scoring function that appears to be able to reliably pick out active ligands from similar decoys that fit in a given conformation of the receptor. Combined, these represent a powerful tool set to embark on new drug-discovery projects aimed at the KOR.

5.6 Methods

5.6.1 Ligand preparation

All actives and decoys came from the Cavassoto GPCR Ligand Library and were generated using LigPrep, which forms a reasonable 3-dimensional structure and protonation state of a ligand.

5.6.2 Glide calculations

For SP and XP calculations, the default box and grid generation options were used. Flexible decking was performed using Glide SP and Glide XP without any constraints.

5.6.3 WaterMap

WaterMap was used to generate the crystal water structure within KOR, MOR and DOR. First the proteins were prepared with the Protein Preparation Wizard in Maestro. WaterMap was run in the default mode; the proteins cocrystalized ligands were used to define the binding site area and were removed during the MD simulations of the water structure.

Chapter 6

Details on the Protein Local Optimization Program

6.1 Loop prediction with the single residue library

The full prediction of a loop in its native or homology model environment involves the generation of thousands of individual loop predictions that are evaluated by the energy function. Each of these loop predictions corresponds to one execution of PLOP. We first outline the methods used for single loop prediction and then describe the multiple execution method for full loop prediction that allows for greater sampling and refinement in regions of space that have previously been identified as promising.

6.1.1 Single loop prediction

Single loop prediction starts with an ab initio construction procedure which extensively searches phase space of possible loop geometries that are confined only by the positions of the loops' flanking helices or beta-strands. An execution of PLOP initiates four stages of loop prediction: a. buildup, b. closure, c. clustering, and d. scoring (27). There are several user-given parameters that alter the course of loop prediction. Below is a sample input file (named input.con), referred to throughout the chapter that gives examples of possible parameter values for loop prediction. If this file is in the same directory as the plop executable, one runs it by typing the command "./plop input.con". The resulting log, models, and rmsd files will contain all the information about the predicted loop.

Input.con

```
datadir /your-plop-directory/data/
file usrtempdir /your-directory/PDBID_het_templates/
outdir /your-working-directory/
logfile /your-working-directory/PDBID_stage_ofac050.log
energy para &
  solvent vsgb2.0 &
load pdb /your-directory/previous-stage/PDBID_Model_1.pdb &
```

```
    het yes &
    opt no &
    sym xtal &
    breaks no range_close_breaks A:64 A:71
load native /your-directory/PDBID.pdb
loop predict A:65 A:70 &
  ofac 0.50 &
sideadd file /your-directory/extrares/PDBID.extra_res.list &
segment no &
  niter 1  &
(  maxcalpha 6.0 &  )
  side1 min gboption 11 minimend sideend  &
  side2 gboption 11 sideend &
  side2 min gboption 1 minimend sideend &
  pdbfile your-directory/current-stage/PDBID_init_ofac050.models &
  rmsdfile your-direcotry/current-stage/PDBID_init_ofac050.rmsd
```

6.1.1.a The buildup stage

This first stage generates an initial set of loop conformations via a dihedral angle search. Residues are added sequentially from both loop stems, and the process terminates at the middle (closure) residue. Thousands of loop halves are generated, and if two meet at the closure residue, these two comprise a loop candidate. To build half-loops, PLOP searches through rotamer libraries of backbone dihedral angles that are representative of the Ramachandran plot. The rotamer libraries were developed by recording the backbone dihedral angles of a large database of high-resolution (<2Å) protein crystal structures. These dihedral (f,j) angles were binned every 5˚, resulting in 5˚ backbone libraries that contain 747 (f,j) combinations for Gly, 215 for Pro, and 866 for the rest of the amino acids. Such high resolution is necessary for accurate loop prediction, but makes a brute force combination of all (f,j) angles impossible. To mitigate this, half-loops are built up first at a course resolution (300˚), which decreases (240˚, 180˚, 120˚, 110˚, 100˚, 90˚ … 10˚, 9˚, 8˚, 7˚, 6˚, 5˚) until a maximum of 10^6 loops have been generated. A quick screening is done to reduce computational load by means of an overlap factor (*ofac*), which is a fractional value we choose; multiplying it by the sum of the atomic radii for a pair of atoms gives

a number which we denote the *ofac* cutoff value. If the distance between the two atoms is less than the cutoff, we consider the atoms as clashing, and the loop structure is discarded. A high *ofac* value leads to greater loop rejection and fewer structures, while a low one allows greater steric clash and more loop candidates.

In input.con, the "loop predict A:65 A:70 &" line calls the main loop buildup subroutines for a loop with residues 65-70; A denotes the protein chain. "ofac 0.50 &" specifies the *ofac* value. The line "breaks no range_close_breaks A:64 A:71" ensures that the residues flanking the loop cannot contain a chain break (missing electron density information). If there is a break in other parts of the crystal structure, PLOP adds an oxygen atom to the C terminal to make COO- and two hydrogens to the N terminal to make NH3+.

6.1.1.b The closure stage

Pairs of half-loops with closure Cα atoms within 0.5Å of each other comprise loop candidates. The positions of the closure Cα are averaged, and the Cβ, Hα and side chain atoms of the closure residue are added to complete the candidate loop. Its N-Cα-C angle is required to be within 25˚ of the ideal value of 111.1˚, and backbone dihedrals are checked against occupied areas of Ramachandran space. A steric clash screening between the two halves and between the closure residue's side chain and the rest of the protein is run to confirm that this fused loop candidate is self-compatible. At this point, there can still be tens of thousands of possible loop candidates. Optimizing and scoring each is prohibitively expensive and redundant, as many will be structurally similar.

6.1.1.c The clustering stage

To reduce the number of loops for energetic scoring, the K-means algorithm is used to cluster loops by RMSD. Initially, the maximum number of clusters allowed is four times the

number of loop residues (4n). Afterwards, any cluster with more than 4 times the median variance is split into three clusters, and the clustering algorithm is rerun, this time the maximum number of clusters being 4n+3. This procedure is applied iteratively, up to a maximum 4n+30 clusters. After the final set of clusters is formed, the loop nearest each cluster center (by RMSD) is sent to the last stage of loop prediction: optimization and scoring.

6.1.1.d Optimization and scoring stage

The final set of loop candidates first undergo side chain optimization, which makes use of 10˚ resolution side chain rotamer libraries. At first, all of the side chains are built onto the fixed backbone in random rotamer states. Then each side chain is optimized one at a time (ie. all rotamers are tried and the lowest energy conformation in context of the rest of the rotamers, which are updated each time, is picked). Convergence of the side chains is achieved when less than 5% of the side chains have a lower possible energy rotamer. All heavy-atom torsion angles between the terminal peptide bonds are sampled, and bond lengths and angles associated with these are initially set to default values. After side chain optimization, a full minimization of both the backbone and side chains of the loops is done to remove any remaining clashes (59). During minimization, bond lengths and angles can change, as well as the positions of nonpolar hydrogens. To score the final set of loops, the energy of each is calculated. The line “solvent vsgb2.0” in input.con calls the most recent energy function contained in PLOP. To include ligands in the energy calculation, the command “het yes” is included in the load pdb section of input.con. “het no” is the default setting. Ions yes/no and water yes/no are other options to include (or not) explicit water molecules or ions in the energy calculations. To include ligands, templates need to be generated for each molecule and put into a directory. The line “file usrtempdir /your-directory/PDBID_het_templates/” provides the path to these template files. The

easiest way create these templates is to use the hetgrp_ffgen scripts licensed through the Schrodinger suite.

If crystal symmetry is known, this information can be used by invoking the "sym xtal" line in the load pdb block of input.con. If crystal symmetry is unknown, or the user does not want to use it then the input file should contain "sym none". PLOP explicitly reconstructs crystal cell units by using the dimensions and space groups reported in PDB files. The simulation consists of one asymmetric unit (which can contain one or more protein chains) and all other atoms from nearby symmetric units that are within 30Å of the target loop. All copies of the asymmetric unit are identical at every stage, meaning that any change in the loop structure or side chain packing is updated simultaneously in each unit.

6.1.2 Full loop predictions

A full loop prediction employs a series of multistage, parallel single loop predictions with varying input parameters (58). To automate the procedure, we use a Perl script, termed Metaplop, which creates the input files for the different stages of a full loop prediction.

6.1.2.a Initial stage *(Init)*

We call the first stage of full loop prediction the Init stage. It typically comprises five single loop predictions executed simultaneously with five different ofacs. Each input file generates many loop candidates, and the five lowest energy nonredundant structures (PDBID_Model_X.pdb) from one are used as the starting point for the next stage of loop prediction. Two structures are nonredundant if the global backbone RMSD is less than a user-input variable; a common value is 0.70Å.

In the *Init* stage, the input PDB is either the native crystal structure or a homology model with the loop region removed. The path to this structure goes in the load pdb line.

6.1.2.b First constrained refinedment Stage *(Ref1)*

In *Ref1*, the 25 PDB files associated with the lowest energy structures found from the *Init* stage are loaded for 25 new PLOP executions in which each structure is subjected to further sampling using a Cartesian constraint of a user-specified value (6Å) on the loop's Ca atoms. The "maxcalpha 6.0" line in input.con is in parentheses, because it only appears in the refinement stages. Upon the completion of this stage, the lowest energy nonredundant models are identified from all generated in both the *Init* and *Ref1* stage, and are passed on to the next stage. The user chooses how many loops enter the next stage, but for long loops, particularly in proteins such as GPCRs, 20 is a good number.

6.1.2.c The fixed stages (*Fix1, Fix2, … ,Fixn)*

The general strategy of the fixed stages relies on the smaller conformational flexibility of shorter loops. In the first fixed stage, *Fix1*, the starting loop structures are passed on from the *Ref1* stage, and one of the terminal residues is held fixed in its starting position. Two subdirectories are formed: *Off1* and *Off2*. In *Off1*, 20 PLOP executions are run based on the 20 starting structures with the left terminal residue is held fixed. If the full loop is 14 residues long, with the starting residue fixed, the prediction is now being done for a 13 residue loop. In *Off2* an equivalent set of PLOP executions occur, except the right terminal residue is fixed. Once *Fix1* is complete, the 20 lowest energy structures (this time comprised of loops built in *Init, Ref1,* and *Fix1* stages) are passed onto *Fix2*, which has three possibilities for keeping two residues fixed: the first two residues, the last two residues, or the first and last residue. In *Fix2*, 20 12 residue loops are predicted for each of these three fixed-residue possibilities. The lowest energy structures from all previous stages and *Fix2* are then passed onto *Fix3*. This process continues up until the last *Fixn* stage, which the user pre-specifies. In the fixed stage PLOP executions (not the

Init or *Ref1* stages) two side chain optimizations are run, one for just the smaller loop fragment, and one for the entire loop. Whichever has the lower energy is used in the ranking of final loop candidates.

6.1.2.d Second constrained refinement stage

This stage is identical to the *Ref1* stage, except this time the constraint on the Ca atoms is typically set to 4.0Å. The lowest energy structure generated from all of the stages is the final loop prediction.

6.2 Loop prediction with the dipeptide library

So far, we have only discussed the use of single residue libraries, ie. (ϕ,ψ)angles of each loop residue are sampled. However, the number of loop conformations rapidly explodes as loop length increases, making both the exhaustive sampling computationally impracticable. Furthermore, even if we could fully sample (ϕ,ψ) space, the longer the long loop, the more combinations of (ϕ,ψ) angles exist that can lead to very similar looking loop conformations.

To mitigate these problems, another sampling algorithm was devised to reduce the number of possible loop conformations, as well as increase the realized effective sampling resolution (71). The key is dipeptide sampling, which differs in the step size in torsion angle phase space. As opposed to (ϕ,ψ) angles, the step size of dipeptide sampling is (ϕ_1,ψ_1,ω,ϕ_2,ψ_2). 400 dipeptide torsion angles libraries, binned at every 5˚, were constructed from 3799 crystal structures which have less than 30% sequence identity and are of <2Å resolution. Average dipeptide libraries only hvae 211 angles, while the most single residue library contain 866 rotamers. This reduction in the number of angles decreases the number of possible loop conformations and also reduces steric clashes between two adjacent residues and their accompanying side chains: the dipeptide libraries carry more information than their single

residue counterparts.

During the buildup procedure in any of the stages of full loop prediction, the half loops can either have an even or odd number of amino acids. If it is an odd number, then the dipeptide libraries are used for all but the residue closest to the closure residue, and the single residue libraries are used to sample its (ϕ,ψ)space.

In input.con, the line "segment no" calls for the single residue libraries, and should be used for short loops. Segment yes calls for the dipeptide libraries.

6.3 Hierarchical loop prediction with surrounding side chain optimization

The methods described so far ignore potential inaccuracies in the surrounding side chains. As discussed previously, the hierarchical loop prediction with surrounding side chain optimization methodology (HLP-SS) addresses this (62). To use this approach, a file ("sideadd file /your-directory/extrares/PDBID.extra_res.list") that contains a list of all extra side chains to be optimized within a user-specified distance (ie. 7.5Å) is created.

6.3.1 Removal of side chains during backbone sampling

As described before, during loop buildup steric clashes between the loop backbone and the rest of the protein are screened. If a side chain occupies space the loop backbone should occupy, this screen can prevent native-like structures from surviving. To avoid this problem, the user can choose to ignore nearby side chains in the screening clash. The biggest problems associated with ignoring surrounding side chains are (1) conformation search space increases significantly and (2) frequently, even in homology models, initial side chain conformations are approximately correct, and thus valuable information that guides loop buildup is being discarded. For the applications to GPCRs, this option was not used.

6.3.2 Simultaneous optimization of side chains in both loop and nearby regions

In the HLP-SS method, iterative optimization of side chain conformations includes the expanded list of side chains. The loop side chains are optimized first, followed by surrounding side chains. This is important, as native like loops built will have all of the side chains in their nearby environment optimized, thus providing a better starting point for refinement in the next stage. It also prevents false positive bias toward non-native like loop structures that stem from incorrect positioning of nearby side chains. This was crucial to GPCRs, as we used this method to sample the nearby lipid head groups of the membrane when necessary.

6.4 Additional sampling methods

The most recent versions of PLOP contain two other important sampling algorithms that greatly improve loop structure predictions.

6.4.1 Prediction of loops containing small helical secondary structure

With this method, loops containing small helices can be accurately predicted. The helix boundaries, coming from either a similar protein or a secondary structure prediction, are imposed on the calculation such that the helical region is sampled only with a special library of dihedral angles that are commonly found in helices across the Protein Database. Details are summarized in (29) and in chapter 4.

6.4.2 Phase space partitioning method

For some very long loops, particularly in non-native (such as an homology model) environments, even more sampling is required to find the low energy basins in the conformational space search. Our newest algorithm involves placing a plane vertically through the loop, and then running two full loop predictions that require that the closure atom be in one of the two hemispheres. We can analogously enforce that candidate loops be built in four phase

space quadrants. In this way, conformational space is even better sampled than in prior by forcing loops to be tried in regions that may not be identified as low energy early on in the full loop prediction. Details are summarized in (31) and in chapter 3.

6.5 VSGB 2.0: the newest energy model incorporated into PLOP

The VSGB2.0 model uses the energy function described by Eq. 6.1 (28):

$$G_{total} = \sum_{bonds} k_b\left(r - r_o\right)^2 + \sum_{angles} k_\theta\left(\theta - \theta_o\right)^2 + \sum_{torsions} \frac{V_n}{2}\left[1 + \cos\left(n\phi - \delta\right)\right] + \sum_{impropers} k_\varphi\left(\varphi - \varphi_o\right)^2$$

$$+ \sum_{electrostatics} \frac{q_i q_j}{r_{ij}\varepsilon_{in(ij)}} + \sum_{VDW}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + G_{sol} + \sum E_{corrections} \qquad \text{Eq. 6.1}$$

It contains the OPLS-AA protein force field bonded and nonbonded terms, as well as a solvation term (G_{sol}) and several physics-based correction terms. The VSGB 2.0 model approximates G_{sol} with an optimized implicit solvent model that is based on the Surface Generalized Born (SGB) model and the variable dielectric (VD) treatment of polarization from protein side chains. The SGB model is an approximation to the Poisson-Boltzmann equation, and the VD treatment improves its accuracy by varying the internal dialectric constants (between 1.0 and 4.0) which approximately takes into account polarization effects. These internal dielectric constants were reoptimized for VSGB2.0.

What really sets apart VSGB 2.0 from its older counterpart, however, is the series of correction terms that are briefly described below. For a more complete mathematical treatment, refer to the work of Li et al (2011).

6.5.1 Hydrogen bonding correction (E_{HB})

To capture a more accurate multipole description of hydrogen bonding, we have derived an empirical functional form that enforces hydrogen bond angles and distances which was fit to

highly accurate, experimental PDB data (135). The E_{HB} term only applies to hydrogen bonding within the protein (not to protein-solvent hydrogen bonding). It is a function of distances and angles between positively charged nitrogen, negatively charged oxygen or polar atoms in side chains and hydrogen atoms.

6.5.2 Self-contact correction

The side chains of Asn, Gln, Ser, and Thr often interact with their own backbone N or O atoms. This interaction depends on the side chain's conformation and the loop back bone secondary structure, making it a more complex scenario than plain hydrogen bonding. This correction is represented by a sum of Gaussian functions that depend on the distance between a polar atom from a side chain and the backbone N or O atom of its own residue.

6.5.3 π-π packing correction

π-π stacking is crucial for stabilizing many biological structures and processes, but separating these interactions from other nonbonded terms (like the Van der Waals interactions) is challenging. VSGB 2.0 employs an explicit π-π packing correction for pairs of amino acid side chains that include both conventional aromatic rings as well as Y-aromatic structures. It should be noted that π-π stacking also occurs in protein backbones, but for the sake of algorithmic simplicity, only side chain π-π stacking is considered. The energetic correction is a function of the distance between the centers of, the horizontal displacement between, and the dihedral angle between aromatic planes.

6.5.4 Hydrophobic term

The hydrophobic correction term rewards contacts between nonpolar heavy atoms and stabilizes hydrophobic contacts. It attempts to accurately model the interaction between hydrophobic surfaces on proteins and ligands with water. The contact parameters are taken from

a protein-ligand docking energy function (81) and have been optimized to reflect hydrophobic interactions in the protein active site. These parameters lead to larger interactions energies for packing hydrophobic side chains into the protein core than the standard approaches which penalize exposed hydrophobic surface area based on the experimental solvation free energies of small alkanes. Specifically, hydration of small alkanes leads to clathrate structures in which the water molecules do not lose hydrogen bonds, but give up some entropy. In contrast, in a protein active site, the placement of water molecules in a hydrophobic "hole" which should be filled by a hydrophobic side chain would in practice lead to the loss of hydrogen bonds for those waters for holes of typical size. Continuum solvation models treat water molecules as infinitesimal dipoles, so the loss of hydrogen bonding based on the physical size of the water molecule compared to the dimensions of a hydrophobic cavity in which it is buried cannot be properly evaluated by such models. Consequently, it is essential to complement the continuum electrostatics approach with a hydrophobic term that takes such effects into account; the efficaciousness of the hydrophobic term that we use, as compared to alternatives, is demonstrated in the work of Li et al (2011), which presents major improvements in loop prediction when the new hydrophobic term is utilized.

Conclusions

From this work, we see a significant advance in modeling G-Protein coupled receptors. First, we have shown vast improvement in predicting the three-dimensional structure of their flexible loop domains within the context of a crystal structure. These advances came from a few key observations that were then translated into the PLOP code: first, we determined that including an explicit membrane into the loop prediction calculations when there are important loop-membrane interactions is essential; second, we increased sampling for the longest loops by focusing efforts separately across different regions of space that the loop could occupy; and third, we improved sampling through means already in PLOP, namely lowering the overlap factor, adding extra fixed stages, and enforcing loop helical regions when there was homology modeling-like reasoning to do so. Second, we extended this work to more challenging loop prediction environments. We created artificial perturbed environments where all of the flexible domains and the stabilizing T4-lysozymes were removed, and side chains near the loops being predicted were no longer in their crystallographic positions. Lastly, we successfully built a full homology model of β2AR based on β1AR. One could say that this is a toy model in the sense that the proteins are highly homologous, and we knew *a priori* that the homology model of β2AR without refinement was very good (other than ICL2). However, for cases where we do not have an experimental structure to compare a model to, the best chance of obtaining a trust-worthy model is to use a template that is as similar to the target protein as possible (similarity metrics include sequence identity or closeness along a photogenic tree). Refinement for these cases must be possible before attempting more ambitious homology modeling.

We also see significant advances in docking ligands into the Kappa opioid receptor. Both Glide SP and Glide XP marvelously failed to separate actives from decoys, while the new

scoring function WScore, coupled with KOR specific terms, produces excellent early enrichment. Furthermore, we have elucidated a new understanding of a main binding mode for this receptor to bind antagonists. It appears that the binding is governed by the formation of a salt bridge and the interaction with a high energy water pair near the 5th and 6th transmembrane helices.

Having reached these new frontiers in GPCR modeling lends itself positively to future research. There are several new initiatives to improve very detailed refinement of a loop in PLOP. One of them, currently termed localsamp, has already been implemented, and allows for a dense mesh of loops to be built (and then scored) around a starting loop structure. This is useful for loops where we have a reasonable starting guess structure that we would like to funnel into a lower energy well by only sampling phi-psi angles that will move the next residue in a loop being predicted to a new position that is close to the starting position. While we have already seen this be useful to a real antibody homology modeling case, the technique is currently hindered by an explosion in loop structures. We have embarked on an extension of the project to cut down the number of loops produced, while still sampling very finely around a starting structure. We believe that this, in addition to all of the work done on PLOP over the last 13 years—including research described in this thesis—will lend itself to major future advances in homology modeling, especially for ambitious cases where there does not exist a highly homologous, already crystallized template.

Lastly, our success with WScore is exciting, and we have reason to believe that we will be able to dock into other GPCRs and get similarly good enrichment. We also believe that we can start investigating alternative binding modes of the KOR and begin a real drug discovery effort with experimental collaborators. It would be thoroughly surprising if, within the next

several years, docking projects like these, will not prove to be successful for other GPCRs with known crystal structures, especially as the modeling community has access to increasing numbers of alternative structures of a single GPCR. It would be unsurprising if lead molecule generation and subsequent optimization could even be done reliably on accurate homology models of GPCRs. Indeed, the future of GPCR modeling seems optimistic.

Is this a surprising conclusion given that I started this thesis by claiming we have just entered the era of structure-based GPCR research? No. As topics get increasingly studied and mature, the likelihood of great breakthroughs goes down, and optimism begins to wane. The modeling of GPCRs, however, is still relatively nascent, and there are a great number of highly validated tools—two of which are described in this thesis—that make this research particularly primed for discovery. And these are discoveries with huge potential to have real impact on the human condition, from the production of new pharmaceuticals to creation of techniques that end up being important in other fields as well. This, coupled with the fundamentally interesting and vast biology to which GPCRs are central, will continue to motivate professors, industrial scientists, postdoctoral fellows, and graduate students alike to keep studying these receptors in coming years.

References

1. Fanelli, F. & De Benedetti, P. G. (2011) *Chemical reviews* **111,** PR438-535.

2. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., & Mee, R. P. (1997) *Journal of computer-aided molecular design* **11,** 425-445.

3. Gohlke, H., Hendlich, M., & Klebe, G. (2000) *Journal of molecular biology* **295,** 337-356.

4. Xue, M., Zheng, M., Xiong, B., Li, Y., Jiang, H., & Shen, J. (2010) *Journal of chemical information and modeling* **50,** 1378-1386.

5. Huang, S. Y., Grinter, S. Z., & Zou, X. (2010) *Physical chemistry chemical physics : PCCP* **12,** 12899-12908.

6. Rao, L., Zhang, I. Y., Guo, W., Feng, L., Meggers, E., & Xu, X. (2013) *Journal of computational chemistry*.

7. Scior, T., Bender, A., Tresadern, G., Medina-Franco, J. L., Martinez-Mayorga, K., Langer, T., Cuanalo-Contreras, K., & Agrafiotis, D. K. (2012) *Journal of chemical information and modeling*.

8. Mandal, S., Moudgil, M., & Mandal, S. K. (2009) *European journal of pharmacology* **625,** 90-100.

9. Repasky, M. P., Murphy, R. B., Banks, J. L., Greenwood, J. R., Tubert-Brohman, I., Bhat, S., & Friesner, R. A. (2012) *Journal of computer-aided molecular design* **26,** 787-799.

10. Repasky, M. P., Shelley, M., & Friesner, R. A. (2007) *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **Chapter 8,** Unit 8 12.

11. Friesner, R. A., Murphy, R. B., Repasky, M. P., Frye, L. L., Greenwood, J. R., Halgren, T. A., Sanschagrin, P. C., & Mainz, D. T. (2006) *Journal of medicinal chemistry* **49,** 6177-6196.

12. Halgren, T. A., Murphy, R. B., Friesner, R. A., Beard, H. S., Frye, L. L., Pollard, W. T., & Banks, J. L. (2004) *Journal of medicinal chemistry* **47,** 1750-1759.

13. Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., *et al.* (2004) *Journal of medicinal chemistry* **47,** 1739-1749.

14. Ewing, T. J., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001) *Journal of computer-aided molecular design* **15,** 411-428.

15. Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996) *Journal of molecular biology* **261,** 470-489.

16. Jones, G., Willett, P., Glen, R. C., Leach, A. R., & Taylor, R. (1997) *Journal of molecular biology* **267,** 727-748.

17. Pierce, K. L., Premont, R. T., & Lefkowitz, R. J. (2002) *Nature reviews. Molecular cell biology* **3,** 639-650.

18. Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008) *Current opinion in structural biology* **18,** 581-586.

19. Almen, M. S., Nordstrom, K. J., Fredriksson, R., & Schioth, H. B. (2009) *BMC biology* **7,** 50.

20. Hagn, F., Etzkorn, M., Raschle, T., & Wagner, G. (2013) *J Am Chem Soc* **135,** 1919-1925.

21. Kristiansen, K. (2004) *Pharmacology & therapeutics* **103,** 21-80.

22. Lappano, R. & Maggiolini, M. (2011) *Nature reviews. Drug discovery* **10,** 47-60.

23. Kolakowski, L. F., Jr. (1994) *Receptors & channels* **2,** 1-7.

24. Zhang, M. & Wang, W. (2003) *Accounts of chemical research* **36,** 530-538.

25. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235-242.

26. Patny, A., Desai, P. V., & Avery, M. A. (2006) *Current medicinal chemistry* **13,** 1667-1691.

27. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., & Friesner, R. A. (2004) *Proteins* **55,** 351-367.

28. Li, J., Abel, R., Zhu, K., Cao, Y., Zhao, S., & Friesner, R. A. (2011) *Proteins* **79,** 2794-2812.

29. Edward B. Miller, C. M., Kai Zhu, Zuwen Zho, Dahlia A. Goldfeld, Richard A. Friesner (2012).

30. Goldfeld, D. A., Zhu, K., Beuming, T., & Friesner, R. A. (2011) *Proceedings of the National Academy of Sciences of the United States of America* **108,** 8275-8280.

31. Goldfeld, D. A., Zhu, K., Beuming, T., & Friesner, R. A. (2013) *Proteins* **81,** 214-228.

32. Wu, H., Wacker, D., Mileni, M., Katritch, V., Han, G. W., Vardy, E., Liu, W., Thompson, A. A., Huang, X. P., Carroll, F. I., *et al.* (2012) *Nature.*

33. Jorgensen, W. L., Maxwell, D. S., & TiradoRives, J. (1996) *J Am Chem Soc* **118,** 11225-11236.

34. Jorgensen, W. L. & Tiradorives, J. (1988) *J*

Am Chem Soc **110,** 1657-1666.

35. Dorsam, R. T. & Gutkind, J. S. (2007) *Nature reviews. Cancer* **7,** 79-94.

36. Kobilka, B. & Schertler, G. F. (2008) *Trends in pharmacological sciences* **29,** 79-83.

37. Mustafi, D. & Palczewski, K. (2009) *Molecular pharmacology* **75,** 1-12.

38. Okada, T., Sugihara, M., Bondar, A. N., Elstner, M., Entel, P., & Buss, V. (2004) *Journal of molecular biology* **342,** 571-583.

39. Murakami, M. & Kouyama, T. (2008) *Nature* **453,** 363-367.

40. Park, J. H., Scheerer, P., Hofmann, K. P., Choe, H. W., & Ernst, O. P. (2008) *Nature* **454,** 183-187.

41. Warne, T., Serrano-Vega, M. J., Baker, J. G., Moukhametzianov, R., Edwards, P. C., Henderson, R., Leslie, A. G., Tate, C. G., & Schertler, G. F. (2008) *Nature* **454,** 486-491.

42. Cherezov, V., Rosenbaum, D. M., Hanson, M. A., Rasmussen, S. G., Thian, F. S., Kobilka, T. S., Choi, H. J., Kuhn, P., Weis, W. I., Kobilka, B. K., *et al.* (2007) *Science* **318,** 1258-1265.

43. Jaakola, V. P., Griffith, M. T., Hanson, M. A., Cherezov, V., Chien, E. Y., Lane, J. R., Ijzerman, A. P., & Stevens, R. C. (2008) *Science* **322,** 1211-1217.

44. Michino, M., Abola, E., Brooks, C. L., 3rd, Dixon, J. S., Moult, J., & Stevens, R. C. (2009) *Nature reviews. Drug discovery* **8,** 455-463.

45. Katritch, V., Rueda, M., Lam, P. C., Yeager, M., & Abagyan, R. (2010) *Proteins* **78,** 197-211.

46. de Graaf, C., Foata, N., Engkvist, O., &

Rognan, D. (2008) *Proteins* **71,** 599-620.

47. Lawson, Z. & Wheatley, M. (2004) *Biochemical Society transactions* **32,** 1048-1050.

48. Klco, J. M., Wiegand, C. B., Narzinski, K., & Baranski, T. J. (2005) *Nature structural & molecular biology* **12,** 320-326.

49. Wong, S. K. (2003) *Neuro-Signals* **12,** 1-12.

50. Burstein, E. S., Spalding, T. A., & Brann, M. R. (1998) *The Journal of biological chemistry* **273,** 24322-24327.

51. Cheung, A. H., Dixon, R. A., Hill, W. S., Sigal, I. S., & Strader, C. D. (1990) *Molecular pharmacology* **37,** 775-779.

52. Chicchi, G. G., Graziano, M. P., Koch, G., Hey, P., Sullivan, K., Vicario, P. P., & Cascieri, M. A. (1997) *The Journal of biological chemistry* **272,** 7765-7769.

53. Fiser, A. & Sali, A. (2003) *Bioinformatics* **19,** 2500-2501.

54. Xiang, Z., Soto, C. S., & Honig, B. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7432-7437.

55. Rohl, C. A., Strauss, C. E., Chivian, D., & Baker, D. (2004) *Proteins* **55,** 656-677.

56. Nikiforovich, G. V., Taylor, C. M., Marshall, G. R., & Baranski, T. J. (2010) *Proteins* **78,** 271-285.

57. Mehler, E. L., Hassan, S. A., Kortagere, S., & Weinstein, H. (2006) *Proteins* **64,** 673-690.

58. Zhu, K., Pincus, D. L., Zhao, S., & Friesner, R. A. (2006) *Proteins* **65,** 438-452.

59. Zhu, K., Shirts, M. R., & Friesner, R. A.

(2007) *J. Chem. Theory Comput.* **3,** 2108-2119.

60. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187-217.

61. Yang, L., Tan, C. H., Hsieh, M. J., Wang, J., Duan, Y., Cieplak, P., Caldwell, J., Kollman, P. A., & Luo, R. (2006) *The journal of physical chemistry. B* **110,** 13166-13176.

62. Sellers, B. D., Zhu, K., Zhao, S., Friesner, R. A., & Jacobson, M. P. (2008) *Proteins* **72,** 959-971.

63. Jones, D. T. (1999) *J. Mol. Biol.* **292,** 195-202.

64. Monge, A., Lathrop, E. J., Gunn, J. R., Shenkin, P. S., & Friesner, R. A. (1995) *Journal of molecular biology* **247,** 995-1012.

65. Parthiban, V., Gromiha, M. M., Abhinandan, M., & Schomburg, D. (2007) *BMC structural biology* **7,** 54.

66. Parthiban, V., Gromiha, M. M., & Schomburg, D. (2006) *Nucleic acids research* **34,** W239-242.

67. Reggio, P. H. (2006) *The AAPS journal* **8,** E322-336.

68. Simms, J., Hall, N. E., Lam, P. H., Miller, L. J., Christopoulos, A., Abagyan, R., & Sexton, P. M. (2009) *Methods Mol Biol* **552,** 97-113.

69. Krieger, E., Nabuurs, S. B., & Vriend, G. (2003) *Methods of biochemical analysis* **44,** 509-523.

70. Sander, C. & Schneider, R. (1991) *Proteins-Structure Function and Genetics* **9,** 56-68.

71. Zhao, S., Zhu, K., Li, J., & Friesner, R. A. (2011) *Proteins* **79,** 2920-2935.

72. Hildebrand, P. W., Goede, A., Bauer, R. A., Gruening, B., Ismer, J., Michalsky, E., & Preissner, R. (2009) *Nucleic acids research* **37,** W571-574.

73. Wu, B., Chien, E. Y., Mol, C. D., Fenalti, G., Liu, W., Katritch, V., Abagyan, R., Brooun, A., Wells, P., Bi, F. C., *et al.* (2010) *Science* **330,** 1066-1071.

74. Chien, E. Y., Liu, W., Zhao, Q., Katritch, V., Han, G. W., Hanson, M. A., Shi, L., Newman, A. H., Javitch, J. A., Cherezov, V., *et al.* (2010) *Science* **330,** 1091-1095.

75. Shimamura, T., Shiroishi, M., Weyand, S., Tsujimoto, H., Winter, G., Katritch, V., Abagyan, R., Cherezov, V., Liu, W., Han, G. W., *et al.* (2011) *Nature* **475,** 65-70.

76. Haga, K., Kruse, A. C., Asada, H., Yurugi-Kobayashi, T., Shiroishi, M., Zhang, C., Weis, W. I., Okada, T., Kobilka, B. K., Haga, T., *et al.* (2012) *Nature* **482,** 547-551.

77. Hanson, M. A., Roth, C. B., Jo, E., Griffith, M. T., Scott, F. L., Reinhart, G., Desale, H., Clemons, B., Cahalan, S. M., Schuerer, S. C., *et al.* (2012) *Science* **335,** 851-855.

78. Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., & Granier, S. (2012) *Nature.*

79. Kruse, A. C., Hu, J., Pan, A. C., Arlow, D. H., Rosenbaum, D. M., Rosemond, E., Green, H. F., Liu, T., Chae, P. S., Dror, R. O., *et al.* (2012) *Nature* **482,** 552-556.

80. Shan, J., Weinstein, H., & Mehler, E. L. (2010) *Biochemistry* **49,** 10691-10701.

81. Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003)

Proteins **52,** 609-623.

82. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I., & Lomize, A. L. (2012) *Nucleic acids research* **40,** D370-376.

83. Lyman, E., Higgs, C., Kim, B., Lupyan, D., Shelley, J. C., Farid, R., & Voth, G. A. (2009) *Structure* **17,** 1660-1668.

84. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., *et al.* (2007) *Bioinformatics* **23,** 2947-2948.

85. (Schrödinger, LLC, New York).

86. Jacobson, M. P., Pincus, D. L., Rapp, C. S., Day, T. J., Honig, B., Shaw, D. E., & Friesner, R. A. (2004) *Proteins: Struct., Funct., Bioinf.* **55,** 351-367.

87. Go, N. & Scheraga, H. A. (1970) *Macromolecules* **3,** 178-187.

88. Palmer, K. A. & Scheraga, H. A. (1991) *J. Comput. Chem.* **12,** 505-526.

89. Moult, J. & James, M. N. G. (1986) *Proteins: Struct., Funct., Bioinf.* **1,** 146-163.

90. Bassolino-Klimas, D. & Bruccoleri, R. E. (1992) *Proteins: Struct., Funct., Bioinf.* **14,** 465-474.

91. DePristo, M. A., de Bakker, P. I. W., Lovell, S. C., & Blundell, T. L. (2003) *Proteins: Struct., Funct., Bioinf.* **51,** 41-55.

92. de Bakker, P. I. W., DePristo, M. A., Burke, D. F., & Blundell, T. L. (2003) *Proteins: Struct., Funct., Bioinf.* **51,** 21-40.

93. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W., & Kollman, P. A. (1995) *J. Am. Chem. Soc.*

117, 5179-5197.

94. Guvench, O., Weiser, J., Shenkin, P., Kolossvary, I., & Still, W. C. (2002) *Journal of computational chemistry* **23,** 214-221.

95. Petrey, D. & Honig, B. (2005) *Mol. Cell* **20,** 811-819.

96. Knight, S., Andersson, I., & Branden, C. I. (1990) *J. Mol. Biol.* **215,** 113-160.

97. Zhu, J., Xie, L., & Honig, B. (2006) *Proteins: Struct., Funct., Bioinf.* **65,** 463-479.

98. Rohl, C. A., Strauss, C. E., Chivian, D., & Baker, D. (2004) *Proteins: Struct., Funct., Bioinf.* **55,** 656-677.

99. Li, X., Jacobson, M. P., & Friesner, R. A. (2004) *Proteins: Struct., Funct., Bioinf.* **55,** 368-382.

100. Wang, G. L. & Dunbrack, R. L. (2003) *Bioinformatics* **19,** 1589-1591.

101. Jones, T. A., Zou, J. Y., Cowan, S. W., & Kjeldgaard, M. (1991) *Acta Crystallogr., Sect. A: Found. Crystallogr.* **47 (Pt 2),** 110-119.

102. Kleywegt, G. J., Harris, M. R., Zou, J. Y., Taylor, T. C., Wahlby, A., & Jones, T. A. (2004) *Acta crystallographica. Section D, Biological crystallography* **60,** 2240-2249.

103. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577-2637.

104. Zhao, S., Zhu, K., Li, J., & Friesner, R. A. (2011) *Proteins: Struct., Funct., Bioinf.* **79,** 2920-2935.

105. Hartigan, J. A. (1975) *Clustering algorithms* (John Wiley, New York).

106. Hartigan, J. A. & Wong, M. A. (1979)

Appl. Stat. **28,** 100-108.

107. Xiang, Z. & Honig, B. (2001) *J. Mol. Biol.* **311,** 421-430.

108. Jacobson, M. P., Friesner, R. A., Xiang, Z., & Honig, B. (2002) *Journal of molecular biology* **320,** 597-608.

109. Ghosh, A., Rapp, C. S., & Friesner, R. A. (1998) *J. Phys. Chem. B* **102,** 10983-10990.

110. Li, X., Jacobson, M. P., Zhu, K., Zhao, S., & Friesner, R. A. (2007) *Proteins: Struct., Funct., Bioinf.* **66,** 824-837.

111. Pollastri, G., Przybylski, D., Rost, B., & Baldi, P. (2002) *Proteins* **47,** 228-235.

112. Sellers, B. D., Zhu, K., Zhao, S., Friesner, R. A., & Jacobson, M. P. (2008) *Proteins: Struct., Funct., Bioinf.* **72,** 959-971.

113. Gouet, P., Courcelle, E., Stuart, D. I., & Metoz, F. (1999) *Bioinformatics* **15,** 305-308.

114. (2011) (Schrodinger, LLC, New York, NY).

115. Koh, I. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., *et al.* (2003) *Nucleic Acids Res.* **31,** 3311-3315.

116. Pirovano, W. & Heringa, J. (2010) *Methods Mol. Biol.* **609,** 327-348.

117. Cendron, L., Berni, R., Folli, C., Ramazzina, I., Percudani, R., & Zanotti, G. (2007) *The Journal of biological chemistry* **282,** 18182-18189.

118. French, J. B. & Ealick, S. E. (2010) *The Journal of biological chemistry* **285,** 35446-35454.

119. Bell, J. A., Ho, K. L., & Farid, R. (2012) *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **68,** 935-952.

120. Sproule, B., Brands, B., Li, S., & Catz-Biro, L. (2009) *Canadian family physician Medecin de famille canadien* **55,** 68-69, 69 e61-65.

121. (2005) in *Medication-Assisted Treatment for Opioid Addiction in Opioid Treatment Programs* (Rockville (MD)).

122. DeLander, G. E., Portoghese, P. S., & Takemori, A. E. (1984) *The Journal of pharmacology and experimental therapeutics* **231,** 91-96.

123. Fields, H. L. (2007) *Regional anesthesia and pain medicine* **32,** 242-246.

124. Nagase, H. & Fujii, H. (2011) *Topics in current chemistry* **299,** 29-62.

125. Vanderah, T. W. (2010) *The Clinical journal of pain* **26 Suppl 10,** S10-15.

126. Wang, Y. H., Sun, J. F., Tao, Y. M., Chi, Z. Q., & Liu, J. G. (2010) *Acta pharmacologica Sinica* **31,** 1065-1070.

127. Bruchas, M. R. & Chavkin, C. (2010) *Psychopharmacology* **210,** 137-147.

128. Chavkin, C. (2011) *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **36,** 369-370.

129. Rives, M. L., Rossillo, M., Liu-Chen, L. Y., & Javitch, J. A. (2012) *The Journal of biological chemistry* **287,** 27050-27054.

130. Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Mathiesen, J. M., Sunahara, R. K., Pardo, L., Weis, W. I., Kobilka, B. K., & Granier, S. (2012) *Nature* **485,** 321-326.

131. Granier, S., Manglik, A., Kruse, A. C., Kobilka, T. S., Thian, F. S., Weis, W. I., & Kobilka, B. K. (2012) *Nature* **485,** 400-404.

132. Thompson, A. A., Liu, W., Chun, E., Katritch, V., Wu, H., Vardy, E., Huang, X. P., Trapella, C., Guerrini, R., Calo, G., *et al.* (2012) *Nature* **485,** 395-399.

133. Fredriksson, R., Lagerstrom, M. C., Lundin, L. G., & Schioth, H. B. (2003) *Molecular pharmacology* **63,** 1256-1272.

134. Waldhoer, M., Bartlett, S. E., & Whistler, J. L. (2004) *Annual review of biochemistry* **73,** 953-990.

135. Morozov, A. V., Kortemme, T., Tsemekhman, K., & Baker, D. (2004) *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6946-6951.